CERTIFIED APPROXIMATION FOR INTERPRETABLE REPRESENTER LANDMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representer explanations rank the training landmarks that most influence a self-supervised representation. At scale, this ranking rests on up to four stacked approximations of the empirical neural tangent kernel (eNTK). These are random output heads, a parameter sketch, landmark sampling and a coefficient fit. Existing analyses bound each approximation separately, but none certifies the top-K set against their combined error. We introduce CAIRN (Certified Approximation for Interpretable Representer laNdmarks), a framework that carries this error through to the ranking. We derive the exact variance of the sketched multi-head eNTK, which matches measurement within 4% where Johnson–Lindenstrauss bounds err by up to $2.5\times$. This yields a high-probability top-K certificate for a fixed coefficient fit, alongside exact residual-trace certificates for discarded spectral mass. An exact product-variance identity separates kernel error from fit variability and identifies when a larger kernel budget can still sharpen a ranking. Stochastic Lanczos Quadrature (SLQ) Ubaru et al. (2017) estimates the effective dimension within 0.72% and guides the landmark budget without dense eigendecomposition. We show that residual mass does not control class coverage, and residual-greedy selection cuts the worst coverage excess of k-means++ from $8.5\times$ to $1.55\times$ ($4\times$ on the sketched eNTK). Cross-view initializers outperform principal-component initialization in five (AUI) to all six (CSI) settings. Against the KREPES Gauss–Newton solver, CAIRN converges 2.5 to $11.3\times$ faster, trails by at most 0.31 points and gains up to 3.14 points on MNIST. Together, these results make the reliability of representer explanations measurable and show where approximation budgets are best spent.

1 INTRODUCTION

The representer theorem states that the minimizer of a regularized empirical risk over a reproducing kernel Hilbert space is a finite kernel expansion over the training points (Kimeldorf & Wahba, 1970; Schölkopf et al., 2001). Representer points (Yeh et al., 2018) and influence functions (Koh & Liang, 2017) carry this idea to deep networks, but only for supervised losses. The eNTK (Jacot et al., 2018; Mohamadi et al., 2023) turns a trained network into a kernel machine, and Nyström landmarks (Williams & Seeger, 2000; Seeger et al., 2003; Alaoui & Mahoney, 2015; Rudi et al., 2015) make these kernels scalable as random projections do for gradient attribution (Park et al., 2023; Schioppa, 2024). KREPES (Zarvandi et al., 2026) combines these concepts for self-supervised learning (SSL), where no labels exist to anchor explanations. It expands the eNTK of a frozen backbone over a few landmarks, fits their coefficients to an SSL objective, and ranks landmarks by coefficient displacement. The ranked list corresponds to the explanation, and in controlled settings its top landmarks share the query's class and expose demographic proxies. Deleting the top-ranked landmarks also collapses model confidence more than deleting random ones (Zarvandi et al., 2026).

We treat reliability as a property of the ranked list, not of the kernel. For a query x_t, KREPES ranks the Nyström landmarks x_ℓ by the sample-specific influence score that measures how far landmark ℓ's coefficients move during the self-supervised fit, and returns the top-K landmarks as the explanation. For large networks, the ranked list rests on four stacked approximations: random output heads, a Subsampled Randomized Hadamard Transform (SRHT) sketch (Ailon & Chazelle, 2006; Tropp, 2011), a Nyström landmark sample and a regularized coefficient fit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior analyses bound each approximation alone against a matrix norm (Gittens & Mahoney, 2016; Woodruff, 2014), but a small kernel error does not imply a stable top-K set. The output heads alone can be coarse, since a single head, as in standard pipelines including KREPES, leaves relative variance $2/r_{\mathrm{tr}}$ in each kernel entry, where r_{tr} is the trace-concentration ratio of the (symmetrized) cross-Jacobian block $J_x J_{x'}^{\top}$ of a single input pair. Two design choices also warrant investigation: landmark pools are judged by residual rather than class coverage, and principal-component initialization is assumed to beat a random start. CAIRN treats the ranked list as the object to certify and composes exact second moments of each approximation up to the ranking. It replaces the analytical KREPES solver with first-order fitting at a fraction of the wall-clock time, and makes the following five contributions.

(i) Exact error model. We derive the SRHT variance in closed form, matching measurement within 4% in experiments, where Johnson–Lindenstrauss (JL) bounds err by up to $2.5\times$. With Gaussian heads, our method yields an exact two-axis decomposition in which part of the parameter variance does not shrink with more heads.

(ii) Top-K certificates. A fixed-fit certificate bounds kernel error, and a repeated-fit certificate adds a coefficient-fit floor that flags when a larger kernel budget can no longer help.

(iii) Residual does not control coverage. Selectors tie on residual yet differ tenfold in coverage, and pivoted Cholesky (Harbrecht et al., 2012) cuts the worst coverage excess from $8.5\times$ to $1.55\times$.

(iv) Contrastive initialization. Our cross-view initializers beat principal-component initialization in 6 (CSI) and 5 (AUI) of 6 settings, with gains of up to 3.8 points on MNIST that exceed the spread across landmark selectors under Barlow Twins and VICReg.

(v) Fast first-order fitting. Spectrum-restricted Lanczos replaces dense landmark eigendecomposition, reproducing the dense eigenbasis to 10^{-13} while reducing working-set memory requirement. With identical initialization, objective, kernel blocks, and landmarks, CAIRN achieves comparable or higher accuracy than KREPES while running 2.5 to $11.3\times$ faster than its Gauss–Newton solve.

CAIRN's novelty lies not in new sketching primitives such as SRHT, Nyström, or Lanczos, but in a framework that propagates approximation uncertainty through the eNTK representer pipeline to the final landmark ranking, including a coefficient-fit component that no kernel budget can remove.

2 Related Work

Self-supervised learning and collapse. Joint-embedding SSL learns by aligning augmented views, either contrastively (Chen et al., 2020) or through asymmetries and statistical penalties that prevent collapse (Grill et al., 2020; Zbontar et al., 2021; Bardes et al., 2022). Jing et al. (2022) showed that these objectives still suffer dimensional collapse. We treat this effect as an input, not a pathology: since output-axis variance scales as $2/(h, r_{\mathrm{tr}}(\bar{M}))$ (Proposition 3.2), a more concentrated eNTK spectrum directly permits fewer output heads than the single-head approximation used by most eNTK pipelines (Mohamadi et al., 2023; Zarvandi et al., 2026).

Representers and attribution. The representer theorem (Schölkopf et al., 2001; Wahba & Wang, 2019) expresses the minimizer of a regularized kernel objective as a finite expansion in the training points, and Belkin et al. (2006) extended such expansions to unlabeled data. Yeh et al. (2018) brought representer points to deep classifiers. Influence functions (Hampel, 1974; Koh & Liang, 2017; Park et al., 2023) attribute predictions through curvature-weighted gradient inner products, and the eNTK (Jacot et al., 2018) makes both tractable for deep networks. KREPES (Zarvandi et al., 2026) combines these ingredients for SSL through a Nyström expansion and a single Gauss–Newton step. Empirical studies show that attributions can be fragile across training runs (Basu et al., 2021; Nguyen et al., 2023). Closest to our goal, Hu et al. (2026) and Tong et al. (2026) show that the random projection and TRAK's approximations largely preserve influence scores and their ranking.

Nyström and sketching. Nyström approximation (Williams & Seeger, 2000) is analyzed through spectral bounds (Gittens & Mahoney, 2016) and computational regularization (Rudi et al., 2015; 2017). Landmarks are chosen by leverage scores (Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018), adaptive rules (Calandriello et al., 2017; Patel et al., 2015) or pivoted Cholesky (Harbrecht et al., 2012). Its one-sidedness $\widetilde{Q} \preceq K_{nn}$ is classical (Horn & Johnson, 2012; Quiñonero-Candela & Rasmussen, 2005). SRHT (Ailon & Chazelle, 2006; Tropp, 2011; Boutsidis

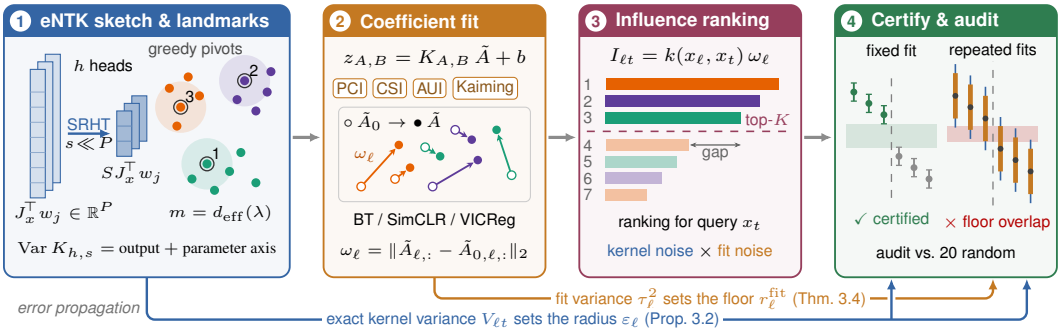

Figure 1: **CAIRN pipeline.** Stage 1 sketches the eNTK and selects landmarks by residual-greedy pivoted Cholesky under a $d_{\text{eff}}(\lambda)$ budget. Stage 2 fits the coefficients from one of four initializers, and Stage 3 ranks landmarks by $I_{\ell t}$. Stage 4 certifies the top-K set, flags fit-limited rankings and audits scores by deletion.

& Gittens, 2013; Balabanov et al., 2023) realizes JL embeddings (Johnson & Lindenstrauss, 1984) in $O(P \log P)$ time, and is analyzed through concentration bounds (Halko et al., 2011; Woodruff, 2014; Drineas et al., 2012). Landmarks may also be sampled uniformly or adaptively (Kumar et al., 2012) or placed at k-means (Arthur & Vassilvitskii, 2007) centers (Zhang et al., 2008). Sketched curvature appears in the Newton sketch (Pilanci & Wainwright, 2017), SCOD (Sharma et al., 2021), low-rank GGN access (Dangel et al., 2023) and sketched Lanczos uncertainty (Miani et al., 2024). Kunstner et al. (2019) warn against conflating curvature surrogates with the Hessian. Our initializers relate to canonical correlation analysis (Hotelling, 1936) and spectral contrastive learning (HaoChen et al., 2021). Prior analyses bound single, supervised approximations; CAIRN composes the second moments of four approximations and certifies the resulting top-K set.

3 Method and Theory

Given a trained network with d outputs and P parameters, a representer explanation expands its eNTK in m landmarks, fits coefficients $\tilde{A}$ from an initialization $\tilde{A}_0$, and ranks the landmarks for a query x_t by

$$I_{\ell t} = k(x_\ell, x_t)\,\omega_\ell, \qquad \omega_\ell = \|\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\|_2, \tag{1}$$

where k is the kernel of the fitted expansion, $\tilde{A}$ is the fitted Nyström coefficient matrix and $\tilde{A}_0$ is its initialization. At scale k is not computed exactly, so for a query x_t we ask whether, with probability at least $1-\delta$, the top-K landmarks under the approximated scores are the top-K under the exact-kernel scores, and, when they cannot be certified, which approximation (heads, sketch, landmarks or fit) is responsible. CAIRN approximates the eNTK with h Gaussian output heads and an SRHT of width s, selects the landmarks under an effective-dimension budget, and fits $\tilde{A}$ from a chosen initialization. We derive exact second moments of the randomized kernel (Section 3.1) and carry them to two top-K certificates (Section 3.2): the first certifies kernel approximation for a given fit, and the second exposes a coefficient-fit floor across repeated fits. Sections 3.3 and 3.4 analyze landmark selection, budget and initialization; proofs are in Section C. The symbols are defined in Section A.

3.1 An exact error model for the sketched kernel

Let $J_x \in \mathbb{R}^{d\times P}$ be a fixed Jacobian, zero-padded so that $P \geq 2$ is a power of two. Let D be diagonal with independent Rademacher entries, H the normalized Walsh–Hadamard matrix, and R a uniform selection of s of the P rows without replacement. The row sample, the signs, and the Gaussian heads are mutually independent, and the sketch and composed kernel are given by:

$$S = \sqrt{P/s}\,RHD, \quad K_{h,s}(x,x') = \frac{1}{h}\sum_{j=1}^{h}\langle SJ_x^\top w_j, SJ_{x'}^\top w_j\rangle, \quad w_j \overset{\text{iid}}{\sim} \mathcal{N}(0, d^{-1}I_d). \tag{2}$$

Its mean is the output-averaged eNTK $k^\star(x,x') = \operatorname{tr}(M)/d$, where $M = J_x J_{x'}^\top$ is the $d \times d$ cross-Jacobian of the pair.

Theorem 3.1 (Exact SRHT variance). *For $1 \le s \le P$ and any $C \in \mathbb{R}^{P\times P}$, the estimate $\mathrm{tr}(S^\top SC)$ is unbiased for $\mathrm{tr}\, C$ and*

$$\mathrm{Var}\big[\mathrm{tr}(S^\top SC)\big] = \tfrac{P-s}{(P-1)s}\,\Phi(C), \quad \Phi(C) = 2\big(\|\bar{C}\|_F^2 - \textstyle\sum_j C_{jj}^2\big), \quad \bar{C} = \tfrac{1}{2}(C + C^\top). \tag{3}$$

For $C = vu^\top$ this is the inner-product estimate $\langle Su, Sv\rangle$, with $\Phi = \|u\|^2\|v\|^2 + \langle u, v\rangle^2 - 2\sum_j u_j^2 v_j^2$.

The identity keeps two terms that JL bounds discard: the finite-population factor and the diagonal correction. The factor $(P-s)/(P-1)$ makes the variance vanish at $s = P$, where the sketch is an exact rotation. The diagonal correction lowers the variance most for coordinate-concentrated inputs, which the Hadamard rotation spreads evenly before sampling. For example, $u = v = e_j$ gives $\Phi = 0$, so the SRHT estimate of $\|e_j\|^2$ is exact at every width s. Figure 2(a,b) confirms the identity to within 0.96–1.03 of measurement, while the JL rate over-predicts by up to $2.5\times$.

Proposition 3.2 (Two-axis decomposition). *Let $C_1 = J_{x'}^\top w_1 w_1^\top J_x$ with mean $\bar{C}_1 = J_{x'}^\top J_x/d$, and let $\bar{M} = \frac{1}{2}(M + M^\top)$ be the symmetric part of M. Under Eq. 2 with one sketch shared by all heads,*

$$\mathrm{Var}(K_{h,s}) = \underbrace{\frac{2\|\bar{M}\|_F^2}{h\,d^2}}_{\textit{output axis}} + \underbrace{\frac{P-s}{(P-1)s}\Big(\Phi(\bar{C}_1) + \frac{\mathbb{E}\Phi(C_1) - \Phi(\bar{C}_1)}{h}\Big)}_{\textit{parameter axis}}, \tag{4}$$

and both bracketed terms are nonnegative because Φ is a positive semidefinite quadratic form. When $\mathrm{tr}\,\bar{M} \neq 0$, the output axis contributes $2/(h\, r_{\mathrm{tr}}(\bar{M}))$ to the relative variance, where $r_{\mathrm{tr}}(B) = \mathrm{tr}(B)^2/\|B\|_F^2$.

The ratio r_{tr} measures trace concentration and differs from the stable rank $\|B\|_F^2/\|B\|_2^2$. For an indefinite cross-kernel, trace cancellation can make r_{tr} small and the relative output variance correspondingly large. Heads shrink the output axis and the $1/h$ share of the parameter axis, but $\Phi(\bar{C}_1)$ does not depend on h. Only a wider sketch removes that term, or an independent sketch per head at h times the sketching cost. On an eNTK block, the unsketched cells match the output axis within 7%, and the sketched excess falls with h as Eq. 4 predicts (Section I.1).

3.2 From kernel error to ranking certificates

We separate two targets: the ranking of a fixed fitted model under kernel approximation, and the population ranking across repeated fits. The same kernel moments enter both certificates, while coefficient-fit variability enters only the second.

Fixed-fit certificate. Let $\mathcal{F}$ contain the data, backbone, landmarks $\mathcal{Z}$, initialization, fitted coefficients and a fixed query x_t, so each ω_ℓ is fixed given $\mathcal{F}$. We fix h, s, K and the error allocation, then draw fresh heads and an SRHT independent of $\mathcal{F}$ and define the exact-kernel scores.

$$I^\star_{\ell t} = k^\star(x_\ell, x_t)\omega_\ell, \widehat{I}_{\ell t} = K^{\mathrm{eval}}_{h,s}(x_\ell, x_t)\omega_\ell \tag{5}$$

Then $\mathbb{E}[\widehat{I}_{\ell t} \mid \mathcal{F}] = I^\star_{\ell t}$ and $\mathrm{Var}(\widehat{I}_{\ell t} \mid \mathcal{F}) = \omega_\ell^2 V_{\ell t}$, where $V_{\ell t}$ is the variance (Eq. 4) for the pair (x_ℓ, x_t). The reference $I^\star_{\ell t}$ keeps the fitted coefficients.

Theorem 3.3 (Conditional top-K certificate). *Assume the protocol above with $1 \le K < m$ and $0 < \delta < 1$, and let $\overline{V}_{\ell t} \ge V_{\ell t}$ be valid upper bounds fixed given $\mathcal{F}$. Choose $\delta_\ell > 0$ with $\sum_{\ell=1}^m \delta_\ell \le \delta$, and set $\varepsilon_\ell = \omega_\ell\sqrt{\overline{V}_{\ell t}/\delta_\ell}$. With probability at least $1-\delta$ given $\mathcal{F}$, all m scores satisfy $|\widehat{I}_{\ell t} - I^\star_{\ell t}| \le \varepsilon_\ell$ simultaneously. Let $\widehat{T} = \mathrm{Top}_K(\widehat{I}_{\cdot t})$, and issue a certificate only when*

$$\min_{\ell\in\widehat{T}}(\widehat{I}_{\ell t} - \varepsilon_\ell) > \max_{j\notin\widehat{T}}(\widehat{I}_{jt} + \varepsilon_j), \tag{6}$$

returning uncertified *otherwise. An incorrect certificate is then issued with probability at most δ, and no independence between landmark scores is required.*

With $\delta_\ell = \delta/m$, an observed gap between the K-th and $(K+1)$-th scores above $2\max_\ell \varepsilon_\ell$ suffices. The exact Φ terms are not needed, since any valid upper bound on $V_{\ell t}$ works, for example

$$V_{\ell t} \le k^\star(x_\ell, x_\ell)\, k^\star(x_t, x_t)\Big[\tfrac{2}{h} + \tfrac{P-s}{(P-1)s}\big(2 + \tfrac{4}{h}\big)\Big], \tag{7}$$

where the two diagonal kernel values must be exact or valid upper bounds. A plug-in sample variance does not give the guarantee, because $\overline{V}_{\ell t}$ must be fixed before the evaluation draw. The guarantee covers one query and one audit; simultaneous or adaptively repeated audits need a further split of δ.

Repeated-fit stability and the coefficient floor. Let $\mathcal{G}$ contain the data, backbone, ordered landmark set $\mathcal{Z}$ and query x_t, and condition every moment and probability below on $\mathcal{G}$. Across independent refits, the displacement $\Omega_\ell = \|\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\|_2$ is random, with mean $\bar{\omega}_\ell$ and finite variance τ_ℓ^2. For an evaluation budget $\mathcal{B} = (h, s)$, we draw fresh heads and an SRHT independently of all coefficient-fitting randomness. Then $K_{\ell t}^{\mathcal{B}}$ and Ω_ℓ are independent, and the kernel factor has mean $\kappa_{\ell t} = k^\star(x_\ell, x_t)$ and variance $V_{\ell t}^{\mathcal{B}}$. The landmarks, fitting protocol, and fit distribution stay fixed as $\mathcal{B}$ varies, so changing m or the training kernel lies outside this comparison. The repeated-fit score, its mean and its exact variance are

$$Z_{\ell t}^{\mathcal{B}} = K_{\ell t}^{\mathcal{B}}\Omega_\ell, \qquad \mu_{\ell t} = \mathbb{E}Z_{\ell t}^{\mathcal{B}} = \kappa_{\ell t}\bar{\omega}_\ell, \tag{8}$$

$$\sigma_{\ell t}^2(\mathcal{B}) := \mathrm{Var}(Z_{\ell t}^{\mathcal{B}}) = V_{\ell t}^{\mathcal{B}}\bar{\omega}_\ell^2 + \kappa_{\ell t}^2\tau_\ell^2 + V_{\ell t}^{\mathcal{B}}\tau_\ell^2, \tag{9}$$

and only the middle term is independent of the kernel budget.

Theorem 3.4 (Coefficient-fit floor for repeated-fit top-K stability)**.** *Under the repeated-fit protocol, assume $1 \le K < m$ and $0 < \delta < 1$, and let $T^\star = \mathrm{Top}_K(\mu_{\cdot t})$ under a fixed tie-breaking rule. Choose $\delta_\ell > 0$ with $\sum_\ell \delta_\ell \le \delta$ before any random draw, keep this allocation fixed across all compared budgets, and define the Chebyshev radii*

$$r_\ell(\mathcal{B}) = \sigma_{\ell t}(\mathcal{B})/\sqrt{\delta_\ell}, \qquad r_\ell^{\mathrm{fit}} = |\kappa_{\ell t}|\tau_\ell/\sqrt{\delta_\ell}. \tag{10}$$

Then $r_\ell(\mathcal{B}) \ge r_\ell^{\mathrm{fit}}$ for every budget, and the separation condition

$$\min_{\ell \in T^\star}\big(\mu_{\ell t} - r_\ell(\mathcal{B})\big) > \max_{j \notin T^\star}\big(\mu_{jt} + r_j(\mathcal{B})\big) \tag{11}$$

implies $\Pr[\mathrm{Top}_K(Z_{\cdot t}^{\mathcal{B}}) = T^\star] \ge 1 - \delta$. Consequently, if the floor intervals overlap at the cut,

$$\min_{\ell \in T^\star}\big(\mu_{\ell t} - r_\ell^{\mathrm{fit}}\big) \le \max_{j \notin T^\star}\big(\mu_{jt} + r_j^{\mathrm{fit}}\big), \tag{12}$$

then no evaluation budget makes Eq. 11 hold, and $r_\ell(\mathcal{B}) \to r_\ell^{\mathrm{fit}}$ whenever $V_{\ell t}^{\mathcal{B}} \to 0$.

Theorem 3.4 limits the certificate rather than the ranking, since overlapping second-moment intervals do not by themselves imply an unstable top-K set. If fitting and evaluation share randomness, then $\mathrm{Cov}(K, \Omega) \neq 0$ in general and Eq. 9 no longer holds as an identity. Over 64 independent realizations on 12 landmarks, the plug-in Eq. 9 has a median measured-to-predicted variance ratio of 0.978 (Section C.1). The kernel factor supplies 83–89% of the predicted variance there, so the kernel budget still governs most of the score spread.

Figure 2(c,d) illustrates both certificates on a synthetic backbone with $n = 64$, $m = 12$, $h = 4$, $d = 8$ and 1864 unpadded parameters. The plug-in Chebyshev threshold is 1.91×10^{-3} and the plug-in floor is 7.01×10^{-4}, against a top-5 gap of 2.51×10^{-5}. The floor is $28\times$ the gap, so in this example a larger kernel budget alone cannot close the separation deficit. The example is diagnostic, since τ_ℓ^2 is a plug-in estimate rather than a finite-sample upper confidence bound. Theorem 3.3 thus controls kernel error for one fixed fit, while Theorem 3.4 controls repeated-fit stability of the population score.

3.3 Landmark selection and budget

Let $\widetilde{K} = GG^\top$ with $G \in \mathbb{R}^{n \times q}$ be the Gram matrix of the selection features, and let $\mathcal{Z}$ be the landmark set. For Eq. 2, one valid feature map is $\psi_{h,s}(x) = h^{-1/2}[(SJ_x^\top w_1)^\top, \ldots, (SJ_x^\top w_h)^\top]^\top \in \mathbb{R}^{hs}$, so that $q = hs$. A surrogate selector uses its own features instead, and in every case the Nyström residual is $R = \widetilde{K} - \widetilde{K}_{n\mathcal{Z}}\widetilde{K}_{\mathcal{Z}\mathcal{Z}}^{+}\widetilde{K}_{\mathcal{Z}n}$.

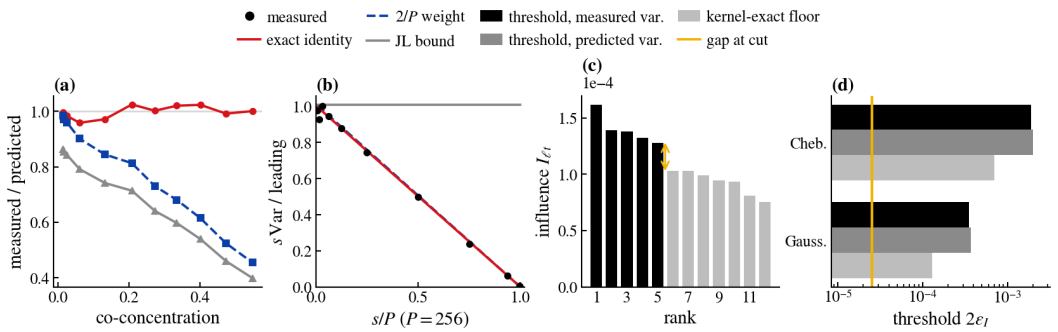

Figure 2: **Error model and certificates. (a)** Measured over predicted SRHT variance ($P = 256$, $s = 32$). **(b)** Finite-population correction. **(c)** Score spectrum at one query. **(d)** Plug-in Chebyshev thresholds; the fit floor is $28\times$ the top-5 gap.

Proposition 3.5 (Greedy pivoted Cholesky)**.** *Let $\delta^{(t)}$ be the residual diagonal after t pivots, and let the next pivot be $j_{t+1} = \arg\max_j \delta_j^{(t)}$. (i) $R_t \succeq 0$ for every realized feature map, and $\sum_j \delta_j^{(t)} = \operatorname{tr} R_t$ holds exactly. (ii) The selected sets are nested, so a single run yields the whole budget curve. (iii) At most $\operatorname{rank}(\widetilde{K}) \le \min(n, q)$ pivots have a positive residual. (iv) The pivot sequence is invariant under the rescaling $\widetilde{K} \mapsto c\widetilde{K}$ for any $c > 0$.*

Part (i) makes the running residual an exact certificate of the kernel mass that the realized features leave unexplained. Part (ii) couples the feature dimension to the budget, since the h-head sketch has at most $\min(n, hs)$ informative pivots. The residual measures total unexplained mass, whereas class coverage is a different functional of the same residual vector.

Proposition 3.6 (Residual mass does not control coverage)**.** *Let class c have index set C_c and trace share $\tau_c = \sum_{j \in C_c} \widetilde{K}_{jj} / \operatorname{tr} \widetilde{K}$. (i) If $\mathcal{Z} \cap C_c = \emptyset$ and $\mathcal{Z}$ spans the features of every point outside C_c, then $\operatorname{tr} R / \operatorname{tr} \widetilde{K} \le \tau_c$ although class c is absent from the pool. (ii) If $|C_c|^{-1} \sum_{j \in C_c} \delta_j^{(t)} > \max_{j \notin C_c} \delta_j^{(t)}$, then the next greedy pivot lies in C_c.*

The KREPES coverage number κ_C ranks the pool by ω_ℓ and counts the landmarks needed to see every class. A class absent from the pool therefore stays uncovered at every rank, however small the residual may be. Part (ii) is a sufficient condition for greedy selection to visit an under-explained class, not a uniform coverage guarantee. For the budget we use the effective dimension $d_{\text{eff}}(\lambda; A) = \operatorname{tr}(A(A + \lambda I)^{-1})$ of a positive semidefinite A (Alaoui & Mahoney, 2015; Rudi et al., 2015).

Lemma 3.7 (Where d_{eff} may be read)**.** *(i) Every principal submatrix $\widetilde{K}_{\mathcal{I}\mathcal{I}}$ and every Nyström approximation $\widetilde{Q} \preceq \widetilde{K}$ satisfies $d_{\text{eff}}(\lambda; \cdot) \le d_{\text{eff}}(\lambda; \widetilde{K})$. (ii) For $\lambda > 0$ and $\widetilde{K} \neq 0$, the effective dimension satisfies $d_{\text{eff}}(\lambda; \widetilde{K}) < \operatorname{rank}(\widetilde{K}) \le \min(n, q)$. (iii) The effective dimension is scale-invariant in the sense that $d_{\text{eff}}(c\lambda; cA) = d_{\text{eff}}(\lambda; A)$ for every $c > 0$.*

Corollary 3.8 (A power-law spectrum has no intrinsic knee)**.** *Any rank-m Nyström approximation leaves a residual with $\operatorname{tr} R \ge \sum_{i>m} \lambda_i(\widetilde{K})$. If $\lambda_i \propto i^{-\alpha}$ with $\alpha > 1$ over the non-saturated range, then $\sum_{i>m} \lambda_i \asymp m^{-(\alpha-1)}$ and $d_{\text{eff}}(\lambda) \asymp \lambda^{-1/\alpha}$ up to the finite-rank cutoff.*

The effective dimension is therefore a scale-aware summary of the realized spectrum, not a certified landmark count for a target ranking error. Reading it from a principal submatrix or a Nyström approximation can only lower it, so such estimates understate the budget the full kernel needs. Under a power-law spectrum, each multiplicative increase in m buys a comparable multiplicative residual reduction, so the data define no unique knee. We therefore use $d_{\text{eff}}(\lambda)$ as a budget heuristic indexed by a user-chosen ridge λ, not as a required landmark count.

3.4 INITIALIZATION

KREPES sets $\tilde{A}_0 = U_k(\Lambda_k + \epsilon I)^{-1/2}$ from the leading eigenpairs of the centered K_{mm}, which we call principal-component initialization (PCI). A Lanczos eigensolver (Lanczos, 1950; Golub & van Loan, 2013) approximates the leading eigenspace from matrix-vector products with $O(mk)$ additional storage.

Proposition 3.9 (Inverse-square-root initialization amplifies spectral error)**.** *Let $(\hat{\lambda}_i, \hat{u}_i)$ be a Ritz pair of a symmetric A, with $\|\hat{u}_i\|_2 = 1$, $\hat{\lambda}_i = \hat{u}_i^\top A \hat{u}_i$ and residual $\rho_i = \|A\hat{u}_i - \hat{\lambda}_i \hat{u}_i\|_2$. Suppose λ_i is the simple eigenvalue closest to $\hat{\lambda}_i$, with gap $\gamma_i = \min_{j \neq i} |\lambda_i - \lambda_j| > 2\rho_i$ and $\lambda_i + \epsilon > 0$. Then, up to sign and to first order in ρ_i, the error of $\hat{u}_i(\hat{\lambda}_i + \epsilon)^{-1/2}$ against $u_i(\lambda_i + \epsilon)^{-1/2}$ is*

$$O\big(\rho_i/(\gamma_i\sqrt{\lambda_i + \epsilon})\big) + O\big(\rho_i/(\lambda_i + \epsilon)^{3/2}\big).$$

The inverse-square-root scaling therefore amplifies errors in small retained eigenvalues and in poorly separated directions. For clustered eigenvalues only the invariant subspace is stable, so the bound should be applied to subspaces rather than individual eigenvectors. This motivates checking the retained Ritz residuals, and it implies that no fixed Lanczos iteration count is universally sufficient. We run $2k + 20$ iterations, capped at m, with full reorthogonalization and verify agreement with the dense initializer separately.

PCI takes its directions from within-view kernel variance, whereas SSL objectives reward agreement between augmented views. We therefore also initialize from cross-view moments of the landmark kernel, using the same relative jitter $\epsilon = 10^{-6} \max_i |\lambda_i|$. Let $K_A, K_B \in \mathbb{R}^{n \times m}$ be the centered kernels of the two views against the landmarks, and define the moments $C_{\text{al}} = \frac{1}{n} K_A^\top K_B$, $C_{\text{all}} = \frac{1}{2n}(K_A^\top K_A + K_B^\top K_B)$ and $B = \frac{1}{2}(C_{\text{al}} + C_{\text{al}}^\top)$. We use

$$\text{CSI}: \ \tilde{A}_0 = U_k\big(|\Lambda_k| + \epsilon I\big)^{-1/2}, \qquad \text{AUI}: \ \tilde{A}_0 = W U_k, \quad W = \big(C_{\text{all}} + \epsilon I\big)^{-1/2}, \tag{13}$$

where CSI takes the leading k eigenpairs of $C_{\text{al}} + C_{\text{al}}^\top$ and AUI takes the leading k eigenvectors of WBW. The absolute value in CSI is needed because $C_{\text{al}} + C_{\text{al}}^\top$ is symmetric but can be indefinite. The columns of the AUI initializer solve the generalized eigenproblem $Bv = \mu\,(C_{\text{all}} + \epsilon I)\,v$ with respect to the pooled within-view covariance. Its eigenvalues are therefore not canonical correlations: canonical correlation analysis (Hotelling, 1936) whitens each view by its own covariance. Both initializers are rescaled to $\|\tilde{A}_0\|_F = \sqrt{k}$, which fixes the overall scale but does not isolate direction from other differences with a Kaiming (He et al., 2015) draw. We ask whether first-order optimization of the SSL loss can replace KREPES's single damped Gauss–Newton step. We test whether it reaches equal accuracy and landmark identifiability in less time, and at what cost.

4 EXPERIMENTS

Setup. CAIRN builds on the KREPES pipeline (Zarvandi et al., 2026) and changes only the components under study (Sections D and E). Stage 1 uses a ResMLP eNTK on Adult and a ConvNet eNTK on MNIST, with $m = 1000$ landmarks per view on Adult and $m = 256$ on MNIST, i.e. two-view pools of $M = 2m = 2000$ and $M = 512$. Stage 2 fits $z = K\tilde{A} + b$ with $k = 256$ under BT (Zbontar et al., 2021), SimCLR (Chen et al., 2020) or VICReg (Bardes et al., 2022), selecting checkpoints by validation accuracy, which uses labels at model-selection time (Section E).

We compare random, k-means++ (KREPES), DPP (Kulesza & Taskar, 2012) and k-center selectors against pivoted Cholesky on a surrogate kernel (PC) or the sketched eNTK (PC-NTK).

Comparison with KREPES. Both methods fit the same encoder from the same PCI initialization, using one damped Gauss–Newton/CG step for KREPES and minibatch Adam for CAIRN. We reimplement the KREPES solver for all three objectives, and every operator matches an explicitly assembled $J^\top H J$ to 10^{-10}. Convergence time includes the KREPES setup. CAIRN keeps its best validation checkpoint, whereas KREPES reports its final CG iterate; Figure 6 shows every KREPES iterate. Identifiability is the percentile of the top-10 deletion effect of the ω_ℓ ranking (Eq. 1) among 30 random deletions (Section F.1).

CAIRN trails KREPES by at most 0.31 points and exceeds it by 1.61 and 3.14 on MNIST SimCLR and VICReg, where KREPES falls below its initialization. It converges 2.5–$11.3\times$ faster in every

Dataset	Objective	Accuracy (%) CAIRN	KREPES	Δ	Time to converge (s) CAIRN	KREPES	Speedup	Identifiability CAIRN / KREPES
Adult	Barlow Twins	**84.10 ± 0.01**	84.01	**+0.09**	**0.994**	10.5	**10.5**×	**83** / 73
Adult	SimCLR	84.08	84.35	-0.27	**2.04**	7.76	**3.8**×	**97** / 93
Adult	VICReg	84.02	84.33	-0.31	**1.57**	17.9	**11.3**×	**83** / 57
MNIST	Barlow Twins	96.30	96.31	-0.01	**1.39**	7.90	**5.7**×	100 / 100
MNIST	SimCLR	**96.10 ± 0.03**	94.49	**+1.61**	**6.10**	20.4	**3.3**×	100 / 100
MNIST	VICReg	**96.13 ± 0.02**	92.99	**+3.14**	**4.63**	11.7	**2.5**×	**100** / 73

Table 1: CAIRN vs. KREPES from the same initialization. Accuracy is k-NN (CAIRN: mean ± s.d., three seeds). Identifiability is the best of three ranking percentiles against 30 random deletions.

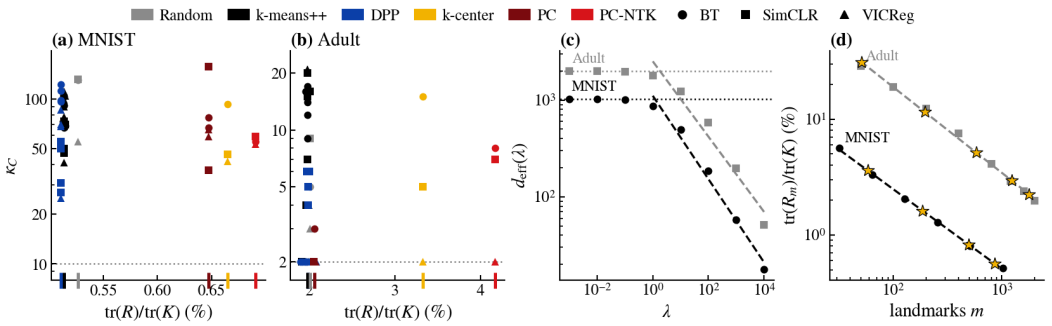

Figure 3: **Coverage and budget. (a,b)** κ_C against Nyström residual; the dotted line is the optimum $\kappa_C = C$. **(c)** $d_{\text{eff}}(\lambda)$ with power-law fits. **(d)** Residual on nested k-means++ prefixes; stars mark $m = d_{\text{eff}}(\lambda)$.

setting, as KREPES pays a setup cost and its CG solve never reaches a tight tolerance (Figures 6 and 7). Its identifiability is never below KREPES, and its top-10 sets are reproducible across seeds, with Jaccard overlap 0.97–1.00. Since loss, initialization, and kernels are shared, these gains are attributable to the solver alone.

Class coverage and landmark budget. Four selectors tie on Nyström residual (0.51–0.65% on MNIST, 1.98–2.07% on Adult), yet residual barely predicts coverage (Spearman $|\rho| \leq 0.26$, $p \geq 0.30$). As Proposition 3.6 predicts, k-means++ on Adult spans $\kappa_C = 2$ to 21 at an unchanged residual (Figure 3a,b). PC has the best mean rank (2.42) and lowest worst-case ratio (1.55 against 8.5 for k-means++), cutting Adult κ_C from 15 to 2 under SimCLR (Table 3). PC adds the point with the largest residual diagonal, so it spreads the pool across regions the kernel has not yet explained. On Adult the PC surrogate has rank 96, so most of its landmarks come from the top-up rule of Section E. The residual decays as a power law, $m^{-0.69}$ on MNIST and $m^{-0.74}$ on Adult ($R^2 \geq 0.996$), so it singles out no budget (Corollary 3.8).For $\lambda \geq 1$, d_{eff} recovers our hand-chosen budgets, with $\lambda = 10$ giving $m = 496$ on MNIST and $m = 1235$ on Adult (Figure 3c,d). The fitted residuals at these budgets lie within 0.03 points of measurement, and SLQ estimates d_{eff} within 0.72%.

Table 2: **Accuracy (%) by initialization.** Only $\tilde{A}_0$ differs; PCI-D/L dense/Lanczos eigensolvers.

	Loss	Kaiming	PCI-D	PCI-L	CSI	AUI
MNIST	BT	90.64	87.15	87.03	90.29	**90.98**
	SimCLR	93.30	96.15	96.24	**97.38**	97.33
	VICReg	**93.23**	90.30	90.21	93.04	93.06
Adult	BT	81.59	81.01	80.95	**81.77**	81.66
	SimCLR	**81.98**	81.29	81.34	81.91	81.86
	VICReg	**81.43**	80.72	80.85	80.87	80.38
Mean rank		2.17	4.17	4.17	**2.00**	2.50

Table 3: **Coverage κ_C by selector** (lower is better; optimum C). Ratios divide by row minimum.

	Loss	Random	k-means++	DPP	k-center	PC-NTK	PC
MNIST	BT	130	67	96	93	**55**	67
	SimCLR	131	50	**31**	46	59	37
	VICReg	55	78	68	**42**	53	65
Adult	BT	5	17	**2**	15	8	3
	SimCLR	9	15	5	5	7	**2**
	VICReg	3	**2**	**2**	**2**	**2**	**2**
Mean rank		4.83	4.58	2.92	3.08	3.17	**2.42**
Mean ratio		2.73	3.61	1.39	2.53	2.11	**1.24**
Worst ratio		4.50	8.50	2.50	7.50	4.00	**1.55**

Initialization of the coefficients. CSI and AUI beat both PCI variants in six and five of six rows, and CSI has the best mean rank (2.00, Table 2). Their largest gains come under SimCLR on MNIST, where they beat Kaiming by 4.1 and 4.0 points. PCI trails Kaiming in five of six rows, while Lanczos matches dense PCI to 4.7×10^{-13} and within 0.13 accuracy points. On MNIST the initializer moves accuracy up to $8.6\times$ more than the selector, which instead governs coverage.

Deletion of influential landmarks. The top-10 landmarks by ω_ℓ pass when deleting them lowers 50-NN confidence more than all 20 random deletions. They pass on Adult SimCLR and VICReg, removing 1.45–$1.58\times$ the damage of the most harmful random deletion. Under a random ranking, two or more passes in six cells has probability 0.030, which is significant at the 5% level. Reselection in coefficient space (Latent-Space Reselection) passes in no cell, and Section D discusses optimizer sensitivity as one limiting factor.

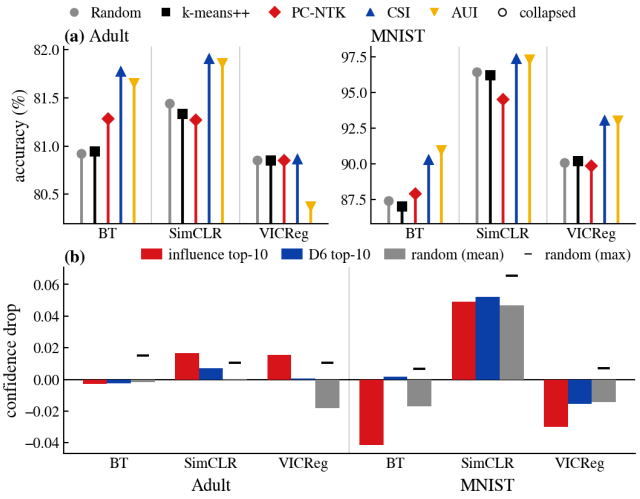

Figure 4: **Accuracy and deletion. (a)** Test accuracy across selectors (PCI-Lanczos fixed) and initializers (k-means++ fixed). **(b)** k-NN confidence drop after deleting the top-10 landmarks by ω_ℓ, Latent-Space Reselection, or 20 random sets (bar: mean; tick: maximum).

5 Conclusion

CAIRN provides a matrix-free framework that propagates randomized eNTK approximation error to representer-landmark rankings and separates kernel-induced instability from coefficient-fit variability. The exact SRHT variance and two-axis decomposition identify how each approximation budget contributes to kernel error, while the resulting high-probability certificate quantifies top-K ranking stability for a fixed fit. Across repeated coefficient fits, the product-variance decomposition further exposes a nonzero variance floor that cannot be removed by increasing the kernel budget; in our worked example, this floor is $28\times$ the top-5 gap, although the estimate is plug-in and finite-sample. Empirically, CAIRN matches KREPES within 0.31 percentage points under matched objectives and initialization, while converging 2.5–$11.3\times$ faster. Pivoted-Cholesky selection achieves accuracy comparable to k-means++ while avoiding its worst class-coverage failures, whereas initializer choice produces larger accuracy changes than selector choice. The confidence-drop study further shows that landmark identifiability is not uniformly detectable.

Future Work. As latents in CAIRN are explicit sums over landmarks, certified landmark rankings enable several key downstream applications, e.g., (i) **Certified Data Governance and Unlearning:** Extending the certificates to eNTK surrogates of large encoders enables data attribution with formal error bars and supports exact top-K deletion for compliance and copyright removal, avoiding approximate retraining; (ii) **Streaming and Federated Coresets:** Extending residual-greedy selection and effective-dimension budgeting to streaming and federated settings could enable dynamic landmark selection without recomputing the full kernel matrix; (iii) **Finite-Sample Bounds for Stochastic Instability:** Developing finite-sample bounds for the coefficient-fit variance floor would provide a general framework for quantifying irreducible instability in stochastic attribution methods.

AI use statement

We used generative AI tools only for proofreading and improving grammar and readability of the manuscript, which falls under the recommended disclosure category.

Ethics statement

Not Applicable.

Reproducibility statement

The source code is submitted as Supplementary Material in a zip folder.

References

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pp. 557–563, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595931341. doi: 10.1145/1132516.1132597. URL https://doi.org/10.1145/1132516.1132597.

Ahmed Alaoui and Michael Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf.

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245. URL https://dl.acm.org/doi/10.5555/1283383.1283494.

Oleg Balabanov, Matthias Beaupère, Laura Grigori, and Victor Lederer. Block subsampled randomized hadamard transform for nyström approximation on distributed architectures. 202:1564–1576, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/balabanov23a.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xHKVVHGDOEk.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7 (85):2399–2434, 2006. URL http://jmlr.org/papers/v7/belkin06a.html.

Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM J. Matrix Anal. Appl.*, 34(3):1301–1340, January 2013. ISSN 0895-4798. doi: 10.1137/120874540. URL https://doi.org/10.1137/120874540.

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Distributed Adaptive Sampling for Kernel Matrix Approximation. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1421–1429. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/calandriello17a.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, July 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Felix Dangel, Lukas Tatzel, and Philipp Hennig. Vivit: Curvature access through the generalized gauss-newton's low-rank structure. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=DzJ7JfPXkE.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (111):3475–3506, 2012. URL http://jmlr.org/papers/v13/drineas12a.html.

Alex Gittens and Michael W. Mahoney. Revisiting the nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17(117):1–65, 2016. URL http://jmlr.org/papers/v17/gittens16a.html.

Gene H. Golub and Charles F. van Loan. *Matrix Computations*. JHU Press, fourth edition, 2013. ISBN 1421407949 9781421407944. URL https://doi.org/10.56021/9781421407944.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962. URL https://doi.org/10.1080/01621459.1974.10482962.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/27debb435021eb68b3965290b5e24c49-Paper.pdf.

Helmut Harbrecht, Michael Peters, and Reinhold Schneider. *Applied Numerical Mathematics*, 62(4):428–440, 2012. ISSN 0168-9274. doi: https://doi.org/10.1016/j.apnum.2011.10.001. URL https://www.sciencedirect.com/science/article/pii/S0168927411001814. Third Chilean Workshop on Numerical Analysis of Partial Differential Equations (WONAPDE 2010).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. URL https://arxiv.org/abs/1502.01852.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012. URL https://doi.org/10.1017/CBO9781139020411.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. URL http://www.jstor.org/stable/2333955.

Pingbang Hu, Yuzheng Hu, Jiaqi W. Ma, and Han Zhao. A unified theory of random projection for influence functions. 2026. URL https://arxiv.org/abs/2602.10449.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. URL http://dx.doi.org/10.1090/conm/026/737400.

George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2239347.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, December 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. volume 13, pp. 981–1006. JMLR.org, April 2012. URL https://jmlr.org/papers/volume13/kumar12a/kumar12a.pdf.

F. Kunstner, P. Hennig, and L. Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 4158–4169. Curran Associates, Inc., December 2019. URL https://dl.acm.org/doi/10.5555/3454287.3454661.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282, October 1950. doi: 10.6028/jres.045.026. URL https://hal.science/hal-01712947.

Marco Miani, Lorenzo Beretta, and Sø ren Hauberg. Sketched lanczos uncertainty score: a low-memory summary of the fisher information. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23123–23154. Curran Associates, Inc., 2024. doi: 10.52202/079017-0728. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29219efdb96e8164c589b4a0124451b7-Paper-Conference.pdf.

Mohamad Amin Mohamadi, Wonho Bae, and Danica J. Sutherland. A fast, well-founded approximation to the empirical neural tangent kernel. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25061–25081. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mohamadi23a.html.

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. A bayesian approach to analysing training data attribution in deep learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64155–64180. Curran Associates, Inc., 2023. doi: 10.52202/075280-2803. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ca774047bc3b46cc81e53ead34cd5d5a-Paper-Conference.pdf.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Raajen Patel, Thomas A. Goldstein, Eva L. Dyer, Azalia Mirhoseini, and Richard G. Baraniuk. oasis: Adaptive column sampling for kernel matrix approximation. 2015. URL https://arxiv.org/abs/1505.05208.

Mert Pilanci and Martin J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017. URL https://doi.org/10.1137/15M1021106.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005. ISSN 1532-4435. URL https://jmlr.org/papers/volume6/quinonero-candela05a/quinonero-candela05a.pdf.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/03e0704b5690a2dee1861dc3ad3316c9-Paper.pdf.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/05546b0e38ab9175cd905eebcc6ebb76-Paper.pdf.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/56584778d5a8ab88d6393cc4cd11e090-Paper.pdf.

Andrea Schioppa. Efficient sketches for training data attribution and studying the loss landscape. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 37692–37735. Curran Associates, Inc., 2024. doi: 10.52202/079017-1190. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4268cf6ba253744befa41ffee5083fa0-Paper-Conference.pdf.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson (eds.), *Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. URL https://doi.org/10.1007/3-540-44581-1_27.

Matthias W. Seeger, Christopher K. I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse gaussian process regression. In Christopher M. Bishop and Brendan J. Frey (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pp. 254–261. PMLR, 03–06 Jan 2003. URL https://proceedings.mlr.press/r4/seeger03a.html.

Apoorva Sharma, Navid Azizan, and Marco Pavone. Sketching curvature for efficient out-of-distribution detection for deep neural networks. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1958–1967. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/sharma21a.html.

Han Tong, Shubhangi Ghosh, Haolin Zou, and Arian Maleki. Imperfect influence, preserved rankings: A theory of trak for data attribution. 2026. URL https://arxiv.org/abs/2602.01312.

Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. 2011. URL https://arxiv.org/abs/1011.1595.

Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $\mathrm{tr}(f(A))$ via stochastic lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017. doi: 10.1137/16M1104974. URL https://doi.org/10.1137/16M1104974.

Grace Wahba and Yuedong Wang. Representer theorem. *Wiley StatsRef: Statistics Reference Online*, 2019. URL https://doi.org/10.1002/9781118445112.stat08200.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, October 2014. ISSN 1551-305X. doi: 10.1561/0400000060. URL https://doi.org/10.1561/0400000060.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8a7129b8f3edd95b7d969dfc2c8e9d9d-Paper.pdf.

Maedeh Zarvandi, Michael Timothy, Theresa Wasserer, and Debarghya Ghoshdastidar. Interpretable self-supervised learning via representer landmarks and nyström approximation. In *Forty-third International Conference on Machine Learning*, volume 306 of *PMLR*, 2026. URL https://openreview.net/forum?id=YMbE22oWDk.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström Low-Rank Approximation and Error Analysis. In *International Conference on Machine Learning*, pp. 1232–1239, 2008. doi: 10.1145/1390156.1390311. URL https://mlanthology.org/icml/2008/zhang2008icml-improved/.

A Appendix index and notation

Appendix index.

- Section A: this index and the notation tables.
- Section B: overview of the four-stage CAIRN architecture (Figure 1).
- Section C: proofs of Theorems 3.1, 3.3 and 3.4 and proposition 3.2, of the conservative variance bound Eq. 7, and of Propositions 3.5, 3.6 and 3.9, lemma 3.7, and corollary 3.8.
- Section D: KREPES and CAIRN algorithms, and how the score $I_{\ell t}$ should be interpreted under first-order optimization.
- Section E: data, selectors, initializers, training and evaluation settings (Table 9).
- Section F: identifiability metrics and per-ranking results (Table 10).
- Section G: optimization dynamics of CAIRN versus KREPES, residual by selector (Table 11) and effective dimension by quadrature.

- Section H: reproduction of KREPES Table 1 (Table 12).
- Section I: empirical verification of the two-axis decomposition (Section I.1) and the product rule (Section C.1).

Notation. The following tables summarize the mathematical notation used throughout the paper, grouped by the section that introduces it (see Tables 4 to 8).

Symbol	Definition
Network, empirical NTK and the sketched kernel	
f_θ	Frozen backbone network with parameters θ; the eNTK is taken at this network and it is never retrained
$x,\ x'$	Generic inputs
d	Number of network outputs
P	Number of network parameters, zero-padded to a power of two for the SRHT (the unpadded count is written out, e.g. 1864)
$J_x \in \mathbb{R}^{d\times P}$	Jacobian of the network output with respect to the parameters at x, zero-padded
$M = J_x J_{x'}^\top,\ \bar{M}$	Cross-Jacobian ($d \times d$) of a pair of inputs; $\bar{M} = \frac{1}{2}(M + M^\top)$ is its symmetric part
$k^\star(x, x')$	Exact output-averaged eNTK, $\mathrm{tr}(M)/d$; the mean of the composed kernel
$k(x, x')$	Kernel entry used in the representer expansion and the influence score Eq. 1
h	Number of Gaussian output heads
w_j	j-th output head, $w_j \overset{\text{iid}}{\sim} \mathcal{N}(0, d^{-1} I_d)$, $j = 1, \dots, h$
s	SRHT sketch width (number of retained rows), $1 \le s \le P$
D	Diagonal Rademacher matrix; its diagonal entries are $\varepsilon_k \in \{\pm 1\}$ in the proof of Theorem 3.1
H	Normalized Walsh–Hadamard matrix ($H_{jk}^2 = 1/P$)
R	Row-selection matrix drawing s of P rows without replacement (in the SRHT only)
$S = \sqrt{P/s}\, RHD$	Subsampled randomized Hadamard transform (SRHT) sketch Eq. 2
$K_{h,s}(x, x')$	Composed kernel estimate: head-averaged inner products of sketched head gradients Eq. 2
$C,\ \bar{C}$	Generic $P \times P$ matrix whose trace $\mathrm{tr}(S^\top S C)$ is estimated (Theorem 3.1); $\bar{C} = \frac{1}{2}(C + C^\top)$
$\Phi(C)$	SRHT variance functional, $2(\|\bar{C}\|_F^2 - \sum_j C_{jj}^2)$ Eq. 3
$\frac{P-s}{(P-1)s}$	Finite-population factor of sampling rows without replacement; vanishes at $s = P$
$u,\ v$	Vectors in the rank-one case $C = vu^\top$, i.e. the inner-product estimate $\langle Su, Sv\rangle$
$a_j,\ T$	In the proof of Theorem 3.1: $a_j = (HDCDH)_{jj}$ and T, the simple random sample of s row indices
$C_1,\ \bar{C}_1$	Single-head matrix $J_{x'}^\top w_1 w_1^\top J_x$ and its mean $J_{x'}^\top J_x / d$ (Proposition 3.2)
$\mathrm{Var}(K_{h,s})$	Output axis $2\|\bar{M}\|_F^2/(hd^2)$ plus parameter axis Eq. 4
$r_{\mathrm{tr}}(B)$	Trace concentration ratio $\mathrm{tr}(B)^2/\|B\|_F^2$; the output axis contributes $2/(h\, r_{\mathrm{tr}}(\bar{M}))$ to relative variance
$\|B\|_F^2/\|B\|_2^2$	Stable rank; contrasted with r_{tr} in Proposition 3.2

Table 4: Notation: network, empirical NTK and the sketched kernel.

B Overview of the CAIRN Architecture

Figure 1 summarizes CAIRN as four stages: matrix-free eNTK approximation and kernel-space landmark selection, self-supervised coefficient fitting, representer influence scoring, and ranking certification. The central design principle is that *the ranked explanation itself is the object to be certified*: each stage exposes its approximation choices and propagates their uncertainty to the final ranking.

Stage 1: Kernel approximation and landmark selection. Starting from a frozen backbone f_θ, CAIRN approximates the eNTK without materializing Jacobians or the full kernel. It projects outputs onto h Gaussian heads $w_a \sim \mathcal{N}(0, I/d)$ and parameters through an SRHT $S = \sqrt{P/s}\, RHD$ of

Symbol	Definition
Influence scores and ranking certificates	
$\mathcal{Z}, m$	Landmark set and its size
x_ℓ	ℓ-th landmark, $\ell = 1, \ldots, m$
x_t	Query (test) point whose landmark ranking is explained
$\tilde{A},\ \tilde{A}_0$	Fitted Nyström coefficient matrix and its initialization; row ℓ is $\tilde{A}_{\ell,:}$
$\Delta\tilde{A}$	Coefficient displacement $\tilde{A} - \tilde{A}_0$
ω_ℓ	Coefficient-displacement weight $\|\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\|_2$ of landmark ℓ Eq. 1
$I_{\ell t}$	Representer influence score $k(x_\ell, x_t)\,\omega_\ell$ Eq. 1
$K, \mathrm{Top}_K(\cdot)$	Size of the top-K cut; the index set of the K largest scores
$\mathcal{F}$	Conditioning information for the fixed-fit certificate: data, backbone, $\mathcal{Z}$, fitted coefficients, initialization and x_t
$K^{\mathrm{eval}}_{h,s}$	Evaluation kernel estimate from fresh heads and a fresh SRHT drawn independently of $\mathcal{F}$
$I^\star_{\ell t},\ \widehat{I}_{\ell t}$	Reference score $k^\star(x_\ell, x_t)\omega_\ell$ and its sketched estimate $K^{\mathrm{eval}}_{h,s}(x_\ell, x_t)\omega_\ell$ Eq. 5
$V_{\ell t},\ \overline{V}_{\ell t}$	Exact variance of the evaluation kernel for the pair (x_ℓ, x_t) (Proposition 3.2), and a valid upper bound on it fixed given $\mathcal{F}$
$\delta,\ \delta_\ell$	Total failure probability; its per-landmark allocation with $\sum_\ell \delta_\ell \le \delta$
ε_ℓ	Conditional Chebyshev radius $\omega_\ell\sqrt{\overline{V}_{\ell t}/\delta_\ell}$ Theorem 3.3
$\widehat{T}$	Observed top-K set $\mathrm{Top}_K(\widehat{I}_{\cdot t})$; certified only under Eq. 6
Ω_ℓ	Random coefficient displacement across independent repeated fits
$\bar{\omega}_\ell,\ \tau_\ell^2$	Mean $\mathbb{E}\Omega_\ell$ and variance $\mathrm{Var}\,\Omega_\ell$ of the displacement
$\mathcal{B} = (h, s)$	Kernel budget
$K^{\mathcal{B}}_{\ell t}$	Evaluation kernel estimate at budget $\mathcal{B}$, independent of Ω_ℓ
$\kappa_{\ell t},\ V^{\mathcal{B}}_{\ell t}$	Mean and variance of $K^{\mathcal{B}}_{\ell t}$; $\kappa_{\ell t}$ does not depend on $\mathcal{B}$
$Z^{\mathcal{B}}_{\ell t}$	Repeated-fit score $K^{\mathcal{B}}_{\ell t}\Omega_\ell$ Eq. 8
$\mu_{\ell t}$	Population score $\kappa_{\ell t}\bar{\omega}_\ell$
$\sigma^2_{\ell t}(\mathcal{B})$	Exact repeated-fit variance $V^{\mathcal{B}}_{\ell t}\bar{\omega}_\ell^2 + \kappa^2_{\ell t}\tau^2_\ell + V^{\mathcal{B}}_{\ell t}\tau^2_\ell$ Eq. 9
$T^\star$	Population top-K set $\mathrm{Top}_K(\mu_{\cdot t})$
$r_\ell(\mathcal{B}),\ r^{\mathrm{fit}}_\ell$	Chebyshev radius $\sigma_{\ell t}(\mathcal{B})/\sqrt{\delta_\ell}$, and the coefficient-fit floor $\|\kappa_{\ell t}\|\tau_\ell/\sqrt{\delta_\ell}$ that it cannot go below Eq. 10

Table 5: Notation: influence scores and the two ranking certificates.

Symbol	Definition
Landmark selection, residual and budget	
$\mathcal{X}$, n	Unlabeled data set and the number of candidate points
M	Landmark pool size (2000 on Adult, 512 on MNIST)
$G \in \mathbb{R}^{n\times q}$, q	Selection feature matrix and its feature dimension
$\widetilde{K} = GG^\top$	Gram matrix of the selection features
$\psi_{h,s}(x)$	h-head sketched feature map $h^{-1/2}[(SJ_x^\top w_1)^\top, \dots, (SJ_x^\top w_h)^\top]^\top \in \mathbb{R}^{hs}$, so $q = hs$
$\widetilde{K}_{n\mathcal{Z}}$, $\widetilde{K}_{\mathcal{Z}\mathcal{Z}}$	Blocks of $\widetilde{K}$ between all points and landmarks, and among landmarks; $^+$ is the pseudoinverse
R	Nyström residual $\widetilde{K} - \widetilde{K}_{n\mathcal{Z}}\widetilde{K}^+_{\mathcal{Z}\mathcal{Z}}\widetilde{K}_{\mathcal{Z}n}$ (R_t after t pivots); $R \succeq 0$
$\mathrm{tr}(R)/\mathrm{tr}(K)$	Relative Nyström residual (Nyström Quality)
$\delta_j^{(t)}$, j_{t+1}	Residual diagonal of point j after t pivots; the next greedy pivot $\arg\max_j \delta_j^{(t)}$
$\preceq$	Loewner (positive semidefinite) order
C, C_c	Number of classes; index set of class c
τ_c	Trace share of class c, $\sum_{j\in C_c} \widetilde{K}_{jj}/\mathrm{tr}\,\widetilde{K}$ (Proposition 3.6)
κ_C	KREPES class coverage number: rank the pool by ω_ℓ and count landmarks until every class appears (Budget Selection); optimum $\kappa_C = C$
λ	User-chosen ridge in the effective dimension
$d_{\mathrm{eff}}(\lambda; A)$	Effective dimension $\mathrm{tr}(A(A+\lambda I)^{-1}) = \sum_i \lambda_i(A)/(\lambda_i(A)+\lambda)$ of a PSD matrix A
$\lambda_i(\cdot)$	i-th eigenvalue, in decreasing order
$\widetilde{K}_{MM}$, $\widetilde{Q}$	A principal submatrix of $\widetilde{K}$, and a Nyström approximation with $\widetilde{Q} \preceq \widetilde{K}$ (Lemma 3.7)
α	Power-law spectral decay exponent, $\lambda_i \propto i^{-\alpha}$ (Corollary 3.8)
K_{nm}, K_{mn}	Data–landmark and landmark–data kernel blocks
K_{mm}, K_{nn}	Landmark–landmark and data–data kernel blocks
K_{pnm}, K_{vm}, K_{tm}	Positive (augmented) view, validation and test kernels against the landmarks
PC, PC-NTK	Greedy pivoted Cholesky on a surrogate kernel, or on the sketched eNTK (the CAIRN selectors)
DPP, k-center	Projection determinantal point process sampling, and farthest-point search, on the sketched eNTK
σ (Alg. 2)	Selector: `rand`, `kmeans`Arthur & Vassilvitskii (2007), `pivchol` (PC), or a sketch-based selector

Table 6: Notation: landmark selection, Nyström residual and landmark budget.

Symbol	Definition
Coefficient initialization	
k	Output (latent) dimension of the coefficient fit ($k = 256$); also the number of retained eigenpairs
$U_k,\ \Lambda_k$	Leading k eigenvectors and eigenvalues of the decomposed matrix
ϵ	Eigenvalue jitter; relative, $10^{-6}\max_i \lvert\lambda_i\rvert$ (KREPES's own PCI uses absolute 10^{-6})
PCI	Principal-component initialization $\tilde{A}_0 = U_k(\Lambda_k + \epsilon I)^{-1/2}$ of the centered K_{mm} (dense or Lanczos eigensolver)
$(\hat{\lambda}_i, \hat{u}_i)$	Approximate (Ritz) eigenpair of a symmetric matrix A (Proposition 3.9)
ρ_i	Ritz residual norm $\lVert A\hat{u}_i - \hat{\lambda}_i\hat{u}_i\rVert_2$
γ_i	Eigengap $\min_{j\neq i}\lvert\lambda_i - \lambda_j\rvert$
$K_A,\ K_B \in \mathbb{R}^{n\times m}$	centered kernels of the two augmented views against the landmarks
C_{al}	Cross-view alignment moment $\frac{1}{n}K_A^\top K_B$
C_{all}	Pooled within-view covariance $\frac{1}{2n}(K_A^\top K_A + K_B^\top K_B)$
B	Symmetrized alignment $\frac{1}{2}(C_{\text{al}} + C_{\text{al}}^\top)$
W	Whitening $(C_{\text{all}} + \epsilon I)^{-1/2}$ used by AUI
CSI	Contrastive spectral initialization, $U_k(\lvert\Lambda_k\rvert + \epsilon I)^{-1/2}$ from $C_{\text{al}} + C_{\text{al}}^\top$ Eq. 13
AUI	Alignment–uniformity initialization WU_k, with U_k the leading eigenvectors of WBW Eq. 13
$\lVert\tilde{A}_0\rVert_F = \sqrt{k}$	Common Frobenius normalization of CSI and AUI

Table 7: Notation: coefficient initialization.

Symbol	Definition
Training, evaluation and identifiability	
$z,\ z_A,\ z_B$	Latent $K\tilde{A} + b$; its two-view versions $K_A\tilde{A} + b$ and $K_B\tilde{A} + b$
$b,\ b_0$	Bias of the kernel encoder (initialized to $0.1 \cdot \mathbf{1}$) and its initial value; z_0 is the latent at $(\tilde{A}_0, b_0)$
$\ell,\ \mathcal{L}$ (Alg. 3)	Self-supervised loss (BT, SimCLR, VICReg, BYOL) and its batch value
$\iota,\ T,\ p$	Initializer, number of epochs and early-stopping patience (Alg. 3)
$\gamma,\ \gamma_0$ (Alg. 1)	KREPES bias and its initial value
λ (Alg. 1)	KREPES Gauss–Newton damping
$g,\ r_b,\ X_b$	KREPES accumulated gradient, batch residual and data batch
$K_{bm}^A,\ K_{bm}^B$	Kernels of the two views of batch b against the landmarks
HVP, jvp, vjp, CG	Hessian-vector product, Jacobian-vector product, vector-Jacobian product, conjugate gradient
$\text{conf}(S)$	k-NN confidence: largest class share among the 50 cosine nearest neighbors of a test point after deleting landmark set S
$\Delta(S)$	Deletion effect $\text{conf}(\emptyset) - \text{conf}(S)$
$\varepsilon = 0.25$	Accuracy-equivalence tolerance in percentage points
ρ	Spearman rank correlation between Nyström Quality and κ_C
$\tilde{x}_j,\ a_j,\ a_j^{(0)}$	In Section F: j-th landmark, j-th row of A, and its initial value
ω_j	Ranking score of landmark j: parameter displacement, contribution or leave-one-out
$S_\omega,\ \Delta_\omega$	Top-K set of ranking ω ($K = 10$) and its deletion effect
$N,\ R_1, \ldots, R_N$	Number of random control sets ($N = 30$) and the sets themselves
π_ω	Percentile: fraction of random deletions with a smaller effect than Δ_ω
p_ω	One-sided permutation p-value $(1 + \#\{t : \Delta(R_t) \geq \Delta_\omega\})/(N+1)$
$S^{(u)}$	Top-K set of the $\lVert A - A_0\rVert$ ranking for training seed u
J	Seed stability: mean pairwise Jaccard overlap of $S^{(1)}, \ldots, S^{(s)}$
Nyström Quality, Budget Selection, Latent-Space Reselection	Evaluation labels: relative residual, coverage κ_C, and reselection in coefficient space

Table 8: Notation: training, evaluation and identifiability.

width s (Equation (2)), yielding $K_{h,s}(x, x') = \frac{1}{h}\sum_a \langle SJ_x^\top w_a, SJ_{x'}^\top w_a\rangle$ (Equation (2)). Its variance decomposes exactly into output-head and parameter-sketch terms (Equation (4)), making the effect of (h, s) explicit. Landmarks are selected directly in kernel space using pivoted Cholesky on the eNTK sketch (PC-NTK), determinantal sampling (DPP), or k-center selection. This avoids input-space surrogates that can minimize a proxy residual while leaving the actual kernel residual unchanged or worse. The landmark budget is chosen from the effective dimension $d_{\text{eff}}(\lambda) = \text{tr}\{K(K + \lambda I)^{-1}\}$ (Lemma 3.7), estimated matrix-free by SLQ. Stage 1 outputs the selected landmarks, the required Nyström kernel blocks, and the exact residual $\text{tr}(K_{nn}) - \text{tr}(K_{mm}^+ K_{nm}^\top K_{nm})$ as an error certificate.

Stage 2: Coefficient initialization and self-supervised fitting. The representation is $z_A = K_{nm}\tilde{A} + b$ and $z_B = K_{pnm}\tilde{A} + b$, with $\tilde{A} \in \mathbb{R}^{M\times k}$. CAIRN initializes $\tilde{A}_0$ using principal-component initialization (PCI) with spectrum-restricted Lanczos, contrastive spectral initialization (CSI), its whitened alignment–uniformity variant (AUI), or Kaiming initialization. The coefficients are then optimized with a self-supervised objective such as SimCLR, Barlow Twins, VICReg, or BYOL. Unlike KREPES's single Gauss–Newton step, CAIRN uses minibatch first-order optimization. Stage 2 returns $\tilde{A}$, b, and the displacement $\Delta\tilde{A} = \tilde{A} - \tilde{A}_0$.

Stage 3: Representer influence. Each landmark receives a coefficient-displacement weight $\omega_\ell = \|\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\|_2$, and a query x_t ranks landmarks using $I_{\ell t} = k(x_\ell, x_t)\omega_\ell$ (Equation (1)). For a fixed kernel expansion, this is an exact algebraic decomposition of the representation's dependence on each landmark. The kernel factor is affected by Stage 1 randomness, whereas ω_ℓ depends on Stage 2 fitting variability; Stage 4 separates these sources.

Stage 4: Certification and empirical audit. CAIRN provides three complementary forms of evidence. *Fixed-fit certification* conditions on ω_ℓ, giving $\text{Var}(\hat{I}_{\ell t} \mid F) = \omega_\ell^2 V_{\ell t}$, with $V_{\ell t}$ determined by Equation (4); concentration bounds and a union bound yield top-K intervals that certify the ranking when they separate at the cut (Theorem 3.3). *Repeated-fit stability* accounts for coefficient variability through the exact product rule (Equation (9)), $\sigma^2(B) = V^B\bar{\omega}^2 + \kappa^2\tau^2 + V^B\tau^2$, where $B = (h, s)$, $V^B = \text{Var}[k]$, $\kappa = \mathbb{E}[k]$, $\bar{\omega} = \mathbb{E}[\omega]$, and $\tau^2 = \text{Var}[\omega]$. Because the term $\kappa^2\tau^2$ persists as B grows, the attainable certificate radius has a nonzero fit floor; when the ranking gap falls below this floor, increasing the kernel budget cannot certify it. *Independent confidence-drop auditing* provides an empirical check: with $z = \text{normalize}(KA + b)$, deleting landmarks S produces $\Delta(S) = \text{conf}_\emptyset - \text{conf}_S$, where confidence is the largest class share among the 50 nearest reference points by cosine similarity. We compare influence-ranked top-10 deletions, 20 random deletions, and Latent-Space Reselected landmarks, requiring the influence drop to exceed the full random-deletion spread.

Output. Each explanation is reported as *certified* when the top-K intervals separate at the chosen δ, *uncertified* when they do not, or *fit-limited* when coefficient variability sets the certification floor. The independent confidence-drop audit is reported alongside these analytic verdicts. Thus, CAIRN propagates the exact second-moment contributions of the head count h, sketch width s, landmark budget m, and coefficient fit to the final influence ranking, explicitly identifying when the available evidence is insufficient to support an explanation.

C PROOFS

Proof of Theorem 3.1. Since H is symmetric and orthogonal, $\text{tr}(S^\top SC) = \frac{P}{s}\sum_{t\in T} a_t$ with $a_j = (HDCDH)_{jj}$ and T a simple random sample of size s. Hence the estimate is unbiased for $\sum_j a_j = \text{tr}\, C$. Given D, sampling without replacement gives variance $\frac{P-s}{(P-1)s}[P\sum_j a_j^2 - (\sum_j a_j)^2]$. Since $\sum_j a_j = \text{tr}\, C$ for every D, the outer variance vanishes. Write $a_j = \sum_{k,l} H_{jk}H_{jl}\varepsilon_k\varepsilon_l C_{kl}$. With $\mathbb{E}[\varepsilon_k\varepsilon_l\varepsilon_{k'}\varepsilon_{l'}]$ nonzero only for paired indices and $H_{jk}^2 = 1/P$, $\mathbb{E}a_j^2 = P^{-2}\big[(\text{tr}\, C)^2 + \sum_{k\neq l}(C_{kl}^2 + C_{kl}C_{lk})\big]$. The sum equals $2(\|\bar{C}\|_F^2 - \sum_k C_{kk}^2) = \Phi(C)$. Summing over j gives $P\,\mathbb{E}\sum_j a_j^2 = (\text{tr}\, C)^2 + \Phi(C)$. For $C = vu^\top$, $\|\bar{C}\|_F^2 = \frac{1}{2}(\|u\|^2\|v\|^2 + \langle u, v\rangle^2)$ and $C_{kk} = u_k v_k$. □

Proof of Proposition 3.2. Condition on the heads. The sketch is unbiased, so $\mathbb{E}_S[K_{h,s} \mid w] = \frac{1}{h}\sum_j w_j^\top M w_j$. For $w \sim \mathcal{N}(0, d^{-1}I)$, $\text{Var}(w^\top Mw) = 2d^{-2}\|\bar{M}\|_F^2$, which gives the output axis.

Given w, $K_{h,s} = \mathrm{tr}(S^\top SC)$ with $C = \frac{1}{h}\sum_j C_j$, so Theorem 3.1 gives the conditional variance. $\Phi(C) = 2\sum_{k\neq l}\bar{C}_{kl}^2$ is a positive semidefinite quadratic form. For independent C_j with mean $\bar{C}_1$, $\mathbb{E}\Phi(\frac{1}{h}\sum_j C_j) = \Phi(\bar{C}_1) + (\mathbb{E}\Phi(C_1) - \Phi(\bar{C}_1))/h$. Jensen gives $\mathbb{E}\Phi(C_1) \geq \Phi(\bar{C}_1) \geq 0$. The law of total variance has no cross-term. Relative form: $\mathbb{E}K_{h,s} = \mathrm{tr}(\bar{M})/d$. □

Proof of Theorem 3.3. Condition on $\mathcal{F}$. Fresh evaluation randomness and Proposition 3.2 give $\mathbb{E}\widehat{I}_{\ell t} = I^\star_{\ell t}$ and $\mathrm{Var}(\widehat{I}_{\ell t}) = \omega_\ell^2 V_{\ell t}$. For positive radii, Chebyshev gives

$$\Pr\left(|\widehat{I}_{\ell t} - I^\star_{\ell t}| > \varepsilon_\ell \mid \mathcal{F}\right) \leq \delta_\ell.$$

A zero radius implies zero variance and exact equality almost surely. A union bound yields simultaneous coverage with probability at least $1-\delta$. On that event, Eq. 6 implies $I^\star_{\ell t} > I^\star_{jt}$ for every $\ell \in \widehat{T}$ and $j \notin \widehat{T}$. Thus an incorrect issued certificate can occur only outside the simultaneous-coverage event. Taking all radii below their maximum gives the stated uniform-gap sufficient condition. □

Proof of Eq. 7. Write $a = \|J_{x_\ell}\|_F^2$ and $b = \|J_{x_t}\|_F^2$. Then $\|\bar{M}\|_F^2 \leq ab$ and $\Phi(\bar{C}_1) \leq 2ab/d^2$. For $u = J_{x_\ell}^\top w$ and $v = J_{x_t}^\top w$, the rank-one formula gives $\Phi(vu^\top) \leq 2\|u\|^2\|v\|^2$. With $A = J_{x_\ell}J_{x_\ell}^\top$ and $B = J_{x_t}J_{x_t}^\top$, Gaussian fourth moments give

$$\mathbb{E}[(w^\top Aw)(w^\top Bw)] = \frac{\mathrm{tr}\,A\,\mathrm{tr}\,B + 2\,\mathrm{tr}(AB)}{d^2} \leq \frac{3ab}{d^2}.$$

Hence $\mathbb{E}\Phi(C_1) \leq 6ab/d^2$. The parameter bracket in Eq. 4 is $(1-1/h)\Phi(\bar{C}_1) + (1/h)\mathbb{E}\Phi(C_1)$, bounded by $(2+4/h)ab/d^2$. Substitute these bounds and use $a/d = k^\star(x_\ell, x_\ell)$ and $b/d = k^\star(x_t, x_t)$. □

Proof of Theorem 3.4. By independence,

$$\mathbb{E}[Z_{\ell t}^{\mathcal{B}}] = \mathbb{E}[K_{\ell t}^{\mathcal{B}}]\mathbb{E}[\Omega_\ell] = \kappa_{\ell t}\bar{\omega}_\ell = \mu_{\ell t}.$$

The variance of a product of independent random variables is

$$\mathrm{Var}[K\Omega] = \mathrm{Var}(K)\mathbb{E}[\Omega]^2 + \mathbb{E}[K]^2\,\mathrm{Var}(\Omega) + \mathrm{Var}(K)\,\mathrm{Var}(\Omega),$$

which gives Eq. 9. All three terms are nonnegative, so

$$\sigma_{\ell t}^2(\mathcal{B}) \geq \kappa_{\ell t}^2\tau_\ell^2$$

and hence $r_\ell(\mathcal{B}) \geq r_\ell^{\mathrm{fit}}$. If $V_{\ell t}^{\mathcal{B}} \to 0$, the first and third terms in Eq. 9 vanish, proving convergence of the radius to the floor.

C.1 Attribution-level Product rule

The score $I_{\ell t} = k(x_\ell, x_t)\,\omega_\ell$ is computed through four approximations: the output heads (h), the parameter sketch (s), the landmark sample ($\mathcal{Z}$, of size m), and the coefficient fit that produces $\tilde{A}$ (which is also where KREPES's curvature solve enters). They do not act on the score symmetrically. Once x_ℓ and x_t are fixed, the heads and the sketch act only on the kernel entry. The landmark sample and the fit act only on the weight, through $\Delta\tilde{A}$. The score is therefore a product of two random factors,

$$Z_{\ell t}^{\mathcal{B}} = K_{\ell t}^{\mathcal{B}}\,\Omega_\ell, \qquad \underbrace{K_{\ell t}^{\mathcal{B}}}_{\text{output and parameter axes}}, \qquad \underbrace{\Omega_\ell}_{\text{sample and fit axes}}, \tag{14}$$

with $\mathbb{E}K_{\ell t}^{\mathcal{B}} = \kappa_{\ell t}$, $\mathrm{Var}\,K_{\ell t}^{\mathcal{B}} = V_{\ell t}^{\mathcal{B}}$ (given exactly by Proposition 3.2), $\mathbb{E}\Omega_\ell = \bar{\omega}_\ell$ and $\mathrm{Var}\,\Omega_\ell = \tau_\ell^2$.

Independence by sample splitting. The two factors are made independent by construction: the heads and sketch used to *evaluate* the kernel row are drawn fresh, independently of all randomness used to *fit* $\tilde{A}$ (landmark draw, initialization, minibatch order and stopping). Under it the variance of the product is an identity, not a delta-method approximation.

Proposition C.1 (Attribution-level product rule)**.** *If $K_{\ell t}^{\mathcal{B}}$ and Ω_ℓ are independent with finite second moments, then $\mathbb{E}Z_{\ell t}^{\mathcal{B}} = \kappa_{\ell t}\bar{\omega}_\ell$ and*

$$\operatorname{Var} Z_{\ell t}^{\mathcal{B}} = \underbrace{V_{\ell t}^{\mathcal{B}}\,\bar{\omega}_\ell^2}_{\textit{kernel factor}} + \underbrace{\kappa_{\ell t}^2\,\tau_\ell^2}_{\textit{weight factor}} + \underbrace{V_{\ell t}^{\mathcal{B}}\,\tau_\ell^2}_{\textit{cross-term}}. \tag{9}$$

Proof. By independence, $\mathbb{E}[(K\Omega)^2] = \mathbb{E}[K^2]\,\mathbb{E}[\Omega^2] = (V+\kappa^2)(\tau^2+\bar{\omega}^2)$. Subtracting $(\mathbb{E}K\Omega)^2 = \kappa^2\bar{\omega}^2$ leaves $V\bar{\omega}^2 + \kappa^2\tau^2 + V\tau^2$. □

What each term measures. The three terms are nonnegative, so their shares of $\operatorname{Var} Z_{\ell t}^{\mathcal{B}}$ sum to one. These shares are what Figure 5(b) plots.

- The *kernel factor* $V_{\ell t}^{\mathcal{B}}\bar{\omega}_\ell^2$ is the kernel error scaled by the typical weight. The head count and sketch width reduce it through Proposition 3.2, and they reduce the cross-term by the same factor.
- The *weight factor* $\kappa_{\ell t}^2\tau_\ell^2$ is the fit error scaled by the exact kernel. It does not depend on $\mathcal{B}$. This is the coefficient-fit floor of Theorem 3.4, where it sets $r_\ell^{\mathrm{fit}} = |\kappa_{\ell t}|\tau_\ell/\sqrt{\delta_\ell}$.
- The *cross-term* $V_{\ell t}^{\mathcal{B}}\tau_\ell^2$ is present even under exact independence. It is small relative to the other two whenever either factor has a small coefficient of variation, since

$$\frac{\text{cross}}{\text{kernel}} = \frac{\tau_\ell^2}{\bar{\omega}_\ell^2} = \mathrm{CV}^2(\Omega_\ell), \qquad \frac{\text{cross}}{\text{weight}} = \frac{V_{\ell t}^{\mathcal{B}}}{\kappa_{\ell t}^2} = \mathrm{CV}^2(K_{\ell t}^{\mathcal{B}}). \tag{15}$$

The first ratio makes the cross-term a direct readout of fit stability. A cross-term that is a few percent of the kernel term means Ω_ℓ varies by only about a sixth of its mean, which is what Figure 5(b) shows.

Why a product-form bound, not a composition bound. A generic composition argument bounds each factor separately and chains the errors. On events $|K-\kappa| \le \varepsilon_K$ and $|\Omega-\bar{\omega}| \le \varepsilon_\Omega$,

$$|Z_{\ell t}^{\mathcal{B}} - \mu_{\ell t}| \le |\bar{\omega}_\ell|\,\varepsilon_K + |\kappa_{\ell t}|\,\varepsilon_\Omega + \varepsilon_K\varepsilon_\Omega, \tag{16}$$

and each event needs its own share of the failure probability. With Chebyshev radii $\varepsilon_K = \sqrt{V_{\ell t}^{\mathcal{B}}/\delta'}$ and $\varepsilon_\Omega = \tau_\ell/\sqrt{\delta'}$ (so $\delta' = \delta_\ell/2$), the right-hand side of Eq. 16 is

$$\frac{\sqrt{V}\,|\bar{\omega}| + |\kappa|\,\tau}{\sqrt{\delta'}} + \frac{\sqrt{V}\,\tau}{\delta'}.$$

The product rule instead applies Chebyshev once to the product. Its radius is $r_\ell(\mathcal{B}) = \sigma_{\ell t}(\mathcal{B})/\sqrt{\delta_\ell}$, where $\sigma_{\ell t} = \sqrt{V\bar{\omega}^2 + \kappa^2\tau^2 + V\tau^2}$. Since $\sqrt{a^2+b^2+c^2} \le a+b+c$, the product radius is never larger than the composition radius. It gains in three ways:

1. the error terms add in quadrature rather than linearly, which is up to $\sqrt{3}$ tighter when the three are comparable;
2. the whole budget δ_ℓ goes to one event instead of being split in two, which gains a further $\sqrt{2}$ on the first-order terms;
3. the cross-term enters at the same $1/\sqrt{\delta_\ell}$ order as the others, instead of the $1/\delta'$ order of the product of two radii.

The price is that the rule controls the variance exactly but not the tails. The radius is a Chebyshev radius and assumes independence of the two factors. If fitting and evaluation share randomness, $\operatorname{Cov}(K,\Omega) \ne 0$ adds $2\bar{\omega}\kappa\operatorname{Cov}(K,\Omega)$ at first order and Eq. 9 no longer holds as an identity.

Chebyshev gives

$$\Pr\big(|Z_{\ell t}^{\mathcal{B}} - \mu_{\ell t}| > r_\ell(\mathcal{B})\big) \le \delta_\ell.$$

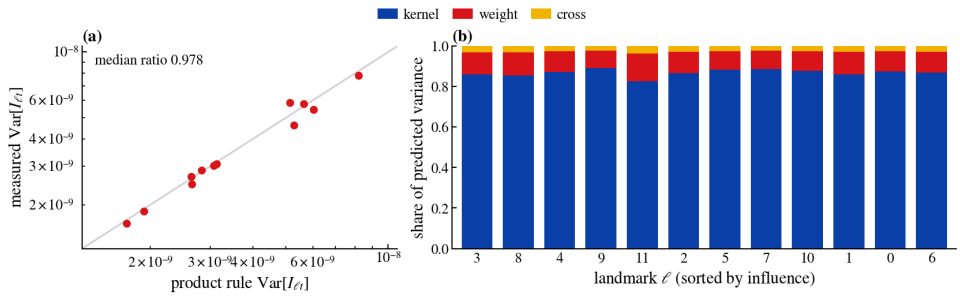

Figure 5: **Equation (9) on all** 12 **landmarks of one test point**, 64 realizations. **(a)** Measured against predicted variance. The median ratio is 0.978; landmarks range from 0.88 to 1.14, within the sampling error of a variance from 64 draws. **(b)** Shares of the predicted variance: kernel term 83 to 89%, weight term 8 to 14%, cross-term a median 2.7%. The cross-term is present under exact independence and is small because ω has a small coefficient of variation.

A union bound therefore yields simultaneous coverage of all m coordinates with probability at least $1-\delta$. On this event, Eq. 11 implies $Z^{\mathcal{B}}_{\ell t} > Z^{\mathcal{B}}_{jt}$ for every $\ell \in T^\star$ and $j \notin T^\star$, so the observed top-K set is $T^\star$.

Finally, because $r_\ell(\mathcal{B}) \geq r^{\text{fit}}_\ell$ for every ℓ, the left-hand side of Eq. 11 is at most $\min_{\ell \in T^\star}(\mu_{\ell t} - r^{\text{fit}}_\ell)$, while its right-hand side is at least $\max_{j \notin T^\star}(\mu_{jt} + r^{\text{fit}}_j)$. Thus Eq. 12 rules out Eq. 11 for every kernel budget. □

Proof of Proposition 3.5. (i) Since $\widetilde{K} = GG^\top$ is a Gram matrix, conditioning on the selected columns gives a generalized Schur complement that is positive semidefinite. Equivalently, R_t is the Gram matrix of the feature vectors after projection onto the orthogonal complement of the span of the selected landmark features. Its diagonal therefore contains the squared projection residuals, and $\sum_j \delta^{(t)}_j = \operatorname{tr} R_t$.

(ii) Once a point is selected, its projection residual is zero, so it is not selected again.

(iii) Each positive pivot adds one linearly independent direction to the selected feature span. Hence there can be at most $\operatorname{rank}(G) = \operatorname{rank}(\widetilde{K}) \leq \min(n, q)$ positive pivots.

(iv) Multiplying $\widetilde{K}$ by $c > 0$ multiplies every residual and every Schur-complement update by the same factor, leaving each $\arg\max_j \delta^{(t)}_j$ unchanged. □

Proof of Proposition 3.6. (i) The residual vanishes off C_c and is at most $\widetilde{K}_{jj}$ on it. (ii) The maximum over C_c is at least the mean, and greedy takes a global maximizer. □

Proof of Lemma 3.7. For a positive semidefinite matrix A with eigenvalues $\lambda_i(A)$,

$$d_{\text{eff}}(\lambda; A) = \sum_i \frac{\lambda_i(A)}{\lambda_i(A) + \lambda}.$$

(i) If $0 \preceq A \preceq B$, then $x \mapsto x/(x+\lambda)$ is operator monotone on the positive semidefinite cone, so $d_{\text{eff}}(\lambda; A) \leq d_{\text{eff}}(\lambda; B)$. For a Nyström approximation this applies directly. For a principal submatrix, Cauchy interlacing and monotonicity of $x/(x+\lambda)$ give the same inequality after summing the corresponding eigenvalues.

(ii) Every nonzero eigenvalue contributes a number strictly between zero and one when $\lambda > 0$. Hence $d_{\text{eff}}(\lambda; \widetilde{K}) < \operatorname{rank}(\widetilde{K})$. Since $\widetilde{K} = GG^\top$ with $G \in \mathbb{R}^{n \times q}$, $\operatorname{rank}(\widetilde{K}) \leq \min(n, q)$.

(iii) Follows from

$$\frac{c\lambda_i(A)}{c\lambda_i(A) + c\lambda} = \frac{\lambda_i(A)}{\lambda_i(A) + \lambda}.$$

□

Proof of Corollary 3.8. $\operatorname{tr} R = \operatorname{tr} \widetilde{K} - \operatorname{tr} \widetilde{Q}$ with $\widetilde{Q} \preceq \widetilde{K}$ of rank at most m, so $\operatorname{tr} \widetilde{Q} \leq \sum_{i \leq m} \lambda_i$ (Horn & Johnson, 2012, §7.4). The rates follow by comparison with $\int_m^\infty t^{-\alpha}\,\mathrm{d}t$ and $\int_1^\infty (1 + \lambda t^\alpha)^{-1}\mathrm{d}t$. □

Proof of Proposition 3.9. For a normalized Ritz pair $(\hat{\lambda}_i, \hat{u}_i)$ with residual ρ_i, standard a posteriori eigenvector perturbation bounds imply that, for a simple eigenvalue separated by gap γ_i, the angular error is of order ρ_i/γ_i. The approximate eigenvalue lies within $O(\rho_i)$ of the target eigenvalue.

Now consider

$$f(\lambda, u) = u(\lambda + \epsilon)^{-1/2}.$$

A first-order perturbation gives

$$\Delta f \approx (\lambda + \epsilon)^{-1/2}\Delta u - \frac{1}{2}u(\lambda + \epsilon)^{-3/2}\Delta\lambda.$$

Using $\|\Delta u\| = O(\rho_i/\gamma_i)$ and $|\Delta\lambda| = O(\rho_i)$ yields

$$\|\Delta f\| = O\left(\frac{\rho_i}{\gamma_i\sqrt{\lambda_i + \epsilon}}\right) + O\left(\frac{\rho_i}{(\lambda_i + \epsilon)^{3/2}}\right).$$

For a clustered eigenvalue the individual eigenvector is not well-conditioned, so the corresponding invariant-subspace perturbation bound should be used instead. □

D ALGORITHMS

Algorithm 1 restates the KREPES procedure (Zarvandi et al., 2026). Algorithms 2 and 3 give the two stages of CAIRN as implemented.

Algorithm 1 KREPES (Zarvandi et al., 2026)

Require: Unlabeled data $\mathcal{X}$, landmark count m, output dimension k, damping λ.
Ensure: Coefficients $\tilde{A} \in \mathbb{R}^{m\times k}$, bias $\gamma \in \mathbb{R}^k$.

Select landmarks $\mathcal{Z} \subset \mathcal{X}$ by k-means++ or leverage-score sampling.
$K_{mm} \approx U_k\Lambda_k U_k^\top$ by dense eigendecomposition; $\tilde{A}_0 \leftarrow U_k\Lambda_k^{-1/2}$, $\gamma_0 \leftarrow \mathbf{0}$.
$g \leftarrow \mathbf{0}$
for each batch $X_b \subset \mathcal{X}$ **do**
    Compute K_{bm}^A, K_{bm}^B for the two views against $\mathcal{Z}$.
    $g \leftarrow g + \nabla_{\tilde{A}}\mathcal{L}_b(K_{bm}^A\tilde{A}_0 + \gamma_0,\ K_{bm}^B\tilde{A}_0 + \gamma_0)$
end for
$\mathrm{HVP}(v) = \sum_b \mathrm{vjp}\big(r_b, \tilde{A}_0, \mathrm{jvp}(r_b, \tilde{A}_0, v)\big) + \lambda v$
$\Delta\tilde{A} \leftarrow \mathrm{CG}(\mathrm{HVP}, -g)$
return $\tilde{A}_0 + \Delta\tilde{A},\ \gamma_0 + \Delta\gamma$

Differences from KREPES and interpretation of the score. KREPES computes a local coefficient update with one damped Gauss–Newton solve at $\tilde{A}_0$. CAIRN instead optimizes $\tilde{A}$ with Adam and early stopping, as Algorithm 3 shows. Consequently,

$$\Delta\tilde{A} = \tilde{A} - \tilde{A}_0$$

is the displacement produced by the specified optimization procedure and can depend on the learning rate, batch order, stopping rule and other optimizer choices.

For a fixed kernel expansion, however, the coefficient displacement still gives the exact algebraic decomposition

$$z(x_t) - z_0(x_t) = \sum_{\ell=1}^{m} k(x_\ell, x_t)\big(\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\big) + (b - b_0).$$

Algorithm 2 CAIRN Stage 1: kernels and landmarks

Require: Unlabeled data $\mathcal{X}$, augmentation, budget m, selector σ.
Ensure: Landmarks $\mathcal{Z}$ and kernels $K_{mm}, K_{nm}, K_{pnm}, K_{vm}, K_{tm}$.
1: **if** $\sigma \in \{\texttt{rand}, \texttt{kmeans}\}$ **then**
2: Select $\mathcal{Z}$ in input space.
3: **else if** $\sigma = \texttt{pivchol}$ **then**
4: Select $\mathcal{Z}$ by greedy pivoted Cholesky on a surrogate kernel (PC).
5: **else**
6: Sketch the eNTK with an SRHT of width s.
7: Select $\mathcal{Z}$ on the sketch by greedy pivoted Cholesky (PC-NTK), projection DPP sampling (DPP) or farthest-point search (k-center).
8: **end if**
9: Repeat selection for the augmented view.
10: Compute K_{mm}, K_{nm}, K_{pnm} (augmented view), K_{vm} (validation) and K_{tm} (test).

Algorithm 3 CAIRN Stage 2: coefficient fit

Require: Kernels from Algorithm 2, loss ℓ, initializer ι, epochs T, patience p.
Ensure: $\tilde{A}$, b and $\tilde{A}_0$.
1: $\tilde{A} \leftarrow$ Kaiming, PCI (dense or Lanczos), CSI or AUI; $\tilde{A}_0 \leftarrow \tilde{A}$; $b \leftarrow 0.1 \cdot \mathbf{1}$.
2: **for** $t = 1$ **to** T **do**
3: **for** each batch $(K_A, K_B) \subset (K_{nm}, K_{pnm})$ **do**
4: $\mathcal{L} \leftarrow \ell(K_A\tilde{A} + b,\ K_B\tilde{A} + b)$
5: Clip $\nabla\mathcal{L}$ to norm 1; take one Adam step on $(\tilde{A}, b)$.
6: **end for**
7: Evaluate validation accuracy; keep the best $(\tilde{A}, b)$; stop after p epochs without gain.
8: **end for**
9: **return** best $(\tilde{A}, b)$ and $\tilde{A}_0$

Thus each landmark contributes an additive vector

$$k(x_\ell, x_t)\big(\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\big)$$

to the change in the fitted representation within this fixed expansion. The scalar score

$$I_{\ell t} = k(x_\ell, x_t)\|\tilde{A}_{\ell,:} - \tilde{A}_{0,\ell,:}\|_2$$

therefore measures a kernel-weighted coefficient displacement. It should not, without an additional argument, be interpreted as the same Gauss–Newton response as KREPES or as a leave-one-out retraining effect.

Optimizer dependence is a separate source of score variability. In particular, Adam's coordinate-wise normalization may change the relative magnitudes of coefficient displacements across landmarks. The sensitivity of the resulting ranking to optimizer seeds, learning rates and stopping rules is therefore an empirical diagnostic, not part of the kernel-approximation certificate.

E Experimental details

Data and kernels. Adult uses a ResMLP eNTK with 3 residual blocks. MNIST uses a ConvNet eNTK. Tabular views are generated by feature masking and Gaussian noise (mask probability 0.1, noise std 0.05). Kernels are cached once per (dataset, selector) and reused by every Stage 2 run.

Selectors. Random uniform and k-means++ select in input space. PC runs greedy pivoted Cholesky on a surrogate kernel. On Adult this kernel has rank 96, so PC selects 96 pivots and fills the remaining 904 of 1000 requested landmarks by the top-up rule. This follows Proposition 3.5 (iii). PC-NTK, DPP and k-center select on an SRHT sketch of the eNTK with $s = 4096$. DPP samples a projection DPP on the sketched features. k-center runs farthest-point search on the sketch.

Initializers. PCI centers K_{mm} and symmetrizes it before the eigensolver. The dense variant uses a float64 eigendecomposition. The Lanczos variant uses float64, full reorthogonalization applied

twice per step and $2k + 20$ iterations. Its random start vector does not advance the global random state, so every initializer leaves the same training trajectory. Both variants use the relative jitter $\epsilon = 10^{-6} \max_i |\lambda_i|$. CSI and AUI center each view, symmetrize every moment before decomposition and use the same relative jitter. KREPES's own PCI (absolute jitter 10^{-6}, no symmetrization, native precision) is kept as a separate arm for reproduction.

Training and evaluation. Stage 2 uses Adam with a cosine schedule and gradient clipping at norm 1. Checkpoints are selected by validation accuracy with patience p. This introduces label information at model-selection time even though the Stage 2 representation loss is self-supervised. The probe is an ℓ_2-regularized linear SVM (liblinear, $C = 1$) with class weights inversely proportional to class frequency. It is fit on normalized validation embeddings and scored by balanced accuracy on the held-out test split. No test label is used for optimizer updates or checkpoint selection. The probe refuses to run when the label sets of the two splits differ. Table 9 lists the remaining settings.

Table 9: Stage 2 settings.

Setting	Value
Output dimension k	256
Epochs / patience	20 / 7
Batch size	512
Learning rate / weight decay	1.25×10^{-2} / 3.0×10^{-5}
BT off-diagonal weight	35
VICReg (λ, μ, ν)	(1.85, 1.61, 0.161)
SimCLR temperature	0.107
Ridge for d_{eff}	10^{-3}
SRHT width s	4096

F IDENTIFIABILITY

Both CAIRN and KREPES learn a kernel encoder whose latent is a sum of per-landmark terms,

$$z(x) \;=\; \sum_{j=1}^{M} k(x, \tilde{x}_j)\, a_j \;+\; b, \tag{17}$$

where $\tilde{x}_j$ is the j-th landmark and a_j is the j-th row of $A \in \mathbb{R}^{M \times d}$. A method is *identifiable* if it can say which landmarks the learned representation relies on.

F.1 IDENTIFIABILITY METRICS

Because of the additive form of Eq. 17, this claim can be tested directly without retraining: remove the landmarks the method ranks highest and measure how much the downstream representation degrades. If the ranking is informative, removing its top landmarks should do more damage than removing the same number of landmarks chosen at random. Table 10 reports exactly this comparison, for three different ways of ranking landmarks.

Landmark rankings. Each ranking assigns a score ω_j to every landmark, where a higher score means the landmark is considered more influential. We consider three scores:

- **Parameter displacement**, $\omega_j = \|a_j - a_j^{(0)}\|_2$: how far the landmark's row of A moved away from the shared initialization A_0 during optimization. The intuition is that landmarks the objective needed to change are the ones it depends on.
- **Contribution**: the magnitude of landmark j's term $k(x, \tilde{x}_j)\, a_j$ in the latent, aggregated over the training set. This measures how much the final representation is built from landmark j, independently of where optimization started.
- **Leave-one-out**: the change in the training-set latents when landmark j alone is removed from Eq. 17.

		$\|A - A_0\|$		Contribution		Leave-one-out		CAIRN seed
Dataset	Objective	CAIRN	KREPES	CAIRN	KREPES	CAIRN	KREPES	stability
Adult	Barlow Twins	7	23	83	70	63	73	1.00
Adult	SimCLR	**97**	70	83	87	90	93	1.00
Adult	VICReg	0	57	83	57	67	33	1.00
MNIST	Barlow Twins	**100**	**100**	**100**	**100**	93	87	0.97
MNIST	SimCLR	**100**	**100**	**100**	0	**100**	0	1.00
MNIST	VICReg	90	73	**100**	30	**100**	27	0.97

Table 10: Landmark identifiability for every ranking. Each cell is the percentile of the ranking's top-landmark deletion effect (drop in k-NN confidence) among 30 random deletions of the same size, median over seeds; **bold**: at least 95%, i.e. one-sided $p < 0.05$. Seed stability: median Jaccard overlap of CAIRN's top landmarks ($\|A - A_0\|$ ranking) across seeds; KREPES is deterministic.

Displacement depends on the optimization trajectory, whereas contribution and leave-one-out depend only on the final encoder.

Deletion effect. Deleting a set S of landmarks means dropping their terms from Eq. 17, for both the reference (training) points and the test points, with all other parameters left unchanged. We measure the effect with the k-NN confidence $\mathrm{conf}(\cdot)$ on the test set: the mean share of each test point's cosine-similarity nearest neighbors in the reference set that belong to its top class. The deletion effect of S is the resulting drop in confidence,

$$\Delta(S) \;=\; \mathrm{conf}(A, b) \;-\; \mathrm{conf}\big(A, b;\, \text{landmarks } S \text{ removed}\big). \tag{18}$$

For a ranking ω, we write $\Delta_\omega = \Delta(S_\omega)$, where S_ω contains the K top-scoring landmarks ($K = 10$ in all experiments).

Random control and percentile. We draw $N = 30$ random sets $R_1, \ldots, R_N$, each of K landmarks sampled uniformly without replacement. The same sets are shared by all rankings and both methods within a run. The table reports the percentile

$$\pi_\omega \;=\; \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}[\Delta(R_t) < \Delta_\omega]\,, \tag{19}$$

that is, the fraction of random deletions that do less damage than deleting the ranking's top landmarks.

Seed stability. Identifiability and reproducibility are separate properties: a ranking can select the same landmarks on every seed and still select the wrong ones. For CAIRN, we therefore also report how consistently its top-K sets agree across training seeds. Let $S^{(1)}, \ldots, S^{(s)}$ be these sets for the $\|A - A_0\|$ ranking. The stability is their mean pairwise Jaccard overlap,

$$J \;=\; \binom{s}{2}^{-1} \sum_{u<v} \frac{|S^{(u)} \cap S^{(v)}|}{|S^{(u)} \cup S^{(v)}|}, \tag{20}$$

and the table reports the median of J across runs. A value of $J = 1$ means every seed selects the same K landmarks. KREPES has no seed dependence, so this column is not defined for it.

F.2 Measuring Identifiability

The per-ranking breakdown in Table 10 shows where the difference comes from. On MNIST, all three CAIRN rankings are significant (≥ 95) for SimCLR and VICReg. KREPES's contribution and leave-one-out rankings fall to 0 and 27–30 there, i.e. no better than random deletions. On Adult, the only significant ranking from either method is CAIRN's $\|A - A_0\|$ on SimCLR (97). Elsewhere on Adult, the two methods are comparable and neither is significant. CAIRN's top landmarks are also stable across seeds, with Jaccard overlap 0.97–1.00.

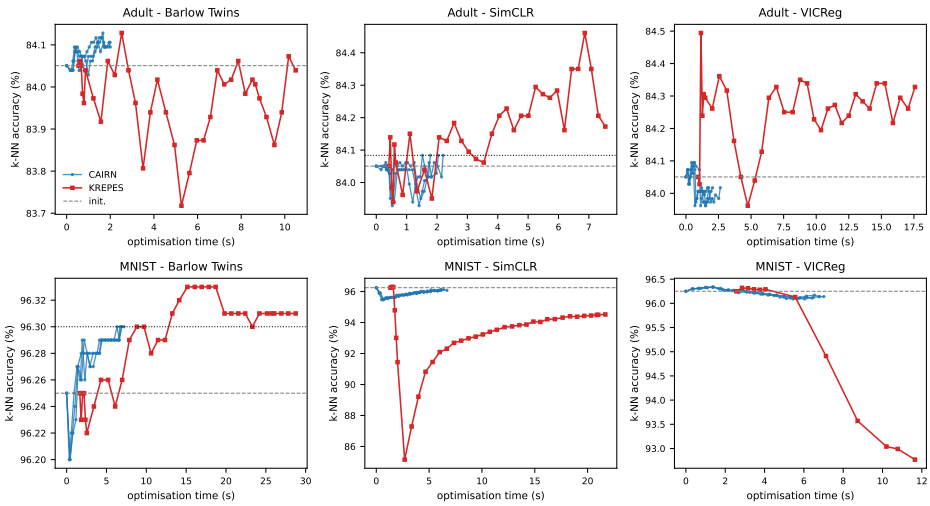

Figure 6: k-NN test accuracy against optimization wall-clock time for CAIRN (blue; one line per seed, three seeds) and KREPES (red; one marker per CG iterate) on Adult (top) and MNIST (bottom). Both methods optimize the same objective from the same initialization, on the same GPU in float32. Dashed grey line: initialization accuracy. Dotted black line: CAIRN's final accuracy. CAIRN settles within about 2.6 s on Adult and 7 s on MNIST. KREPES first incurs a setup cost (gradient pass and preconditioner) before its first iterate, then oscillates on Adult and degrades well below the initialization on MNIST SimCLR and VICReg.

G Additional Results

G.1 Optimization Dynamics

Figure 6 plots k-NN accuracy against wall-clock time.

CAIRN's first update arrives after about 0.2 s. On Adult, its trajectories remain in a narrow band (83.93–84.13%) and finish by 2–2.6 s. KREPES first has to pay a fixed setup cost (gradient pass and preconditioner) of 0.4–2.6 s before its first CG iterate. Its trajectory then oscillates without settling: on Adult Barlow Twins it spans 83.72–84.13%, and on Adult VICReg it spans 83.96–84.49% across 17.6 s. As a result, its final accuracy depends on where CG is stopped. On MNIST, the dynamics differ in kind. On SimCLR, KREPES drops to 85.2% at 2.7 s and recovers only to 94.5% after 21.7 s. On VICReg, it declines monotonically to 92.8%. CAIRN's worst transient is smaller: a dip to 95.5% early on SimCLR, from which it recovers to 96.1%.

Figure 7 explains this behavior: the KREPES linear system is never solved accurately.

On Adult VICReg and MNIST Barlow Twins, it reaches 0.03 and 0.015. SimCLR reaches the lowest residuals ($\approx 10^{-2}$). However, on both datasets its residual oscillates non-monotonically by up to an order of magnitude from one iteration to the next. This pattern is consistent with an ill-conditioned or indefinite system under float32. On MNIST VICReg, CG runs for only 20 iterations and stalls at 0.037. Because each accuracy point in Figure 6 comes from an inexact solve, KREPES produces the noisy and sometimes degrading trajectories shown there. CAIRN avoids the linear solve entirely, which removes both the setup cost and the dependence on CG convergence.

Residual by selector. Table 11 lists Nyström Quality for every selector. The residual depends only on the landmark set, so it is identical across losses and initializations.

Effective dimension by quadrature. SLQ estimates d_{eff} from matrix-vector products alone. Across all 12 (dataset, selector) cells it agrees with the exact value within 0.72%, and within 0.04% in 12 of them. The largest error occurs for PC on Adult, whose kernel is the rank-deficient one noted in Section E.

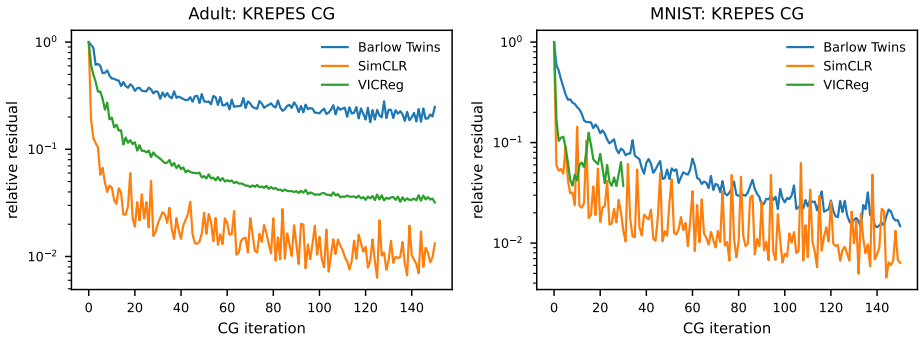

Figure 7: Relative residual of KREPES's preconditioned conjugate-gradient solve against CG iteration, on a log scale, for Adult (left) and MNIST (right). No objective converges to a tight tolerance within 150 iterations. Adult Barlow Twins plateaus near 0.2. SimCLR reaches about 10^{-2} but oscillates by up to an order of magnitude between iterations. MNIST VICReg stops after 20 iterations at about 0.04. The inexact solves account for the non-monotone KREPES trajectories in Figure 6.

Table 11: Relative Nyström residual $\mathrm{tr}(R)/\mathrm{tr}(K)$ (%) by selector.

	Random	k-means++	DPP	k-center	PC-NTK	PC
MNIST	0.526	0.514	**0.511**	0.665	0.691	0.648
Adult	2.012	**1.981**	1.988	3.327	4.168	2.065

H Reproduction of KREPES Table 1

Table 12 reports accuracy and κ_C with Kaiming initialization and k-means++ landmarks. On Adult and MNIST accuracy is nearly insensitive to the loss. The three losses span 0.55 points on Adult and 2.66 on MNIST. κ_C instead varies by an order of magnitude (2 to 21 on Adult, 41 to 91 on MNIST) with no matching movement in accuracy.

Table 12: Test accuracy and κ_C, Kaiming initialization, k-means++ landmarks. κ_C is not comparable across class counts. *Collapsed: top two classes take $\geq 75\%$ of predictions.

Dataset	Loss	Acc (%)	κ_C	Classes seen
Adult	BT	81.59	9	2/2
	SimCLR	81.98	2	2/2
	VICReg	81.43	21	2/2
MNIST	BT	90.64	91	10/10
	SimCLR	93.30	73	10/10
	VICReg	93.23	41	10/10

I Verification of the error model

I.1 Two-axis decomposition

The unsketched cells isolate the output axis and confirm it within 7%. The sweep did not retain the per-head vectors, so the Φ terms of Eq. 4 cannot be evaluated on it exactly. The measured excess over the output term decreases with h at every sketch width, as Eq. 4 predicts. At $s = 20$ it is 0.273, 0.064 and 0.027 for $h = 1, 4, 16$, while the surrogate predicts 0.16 for all three.

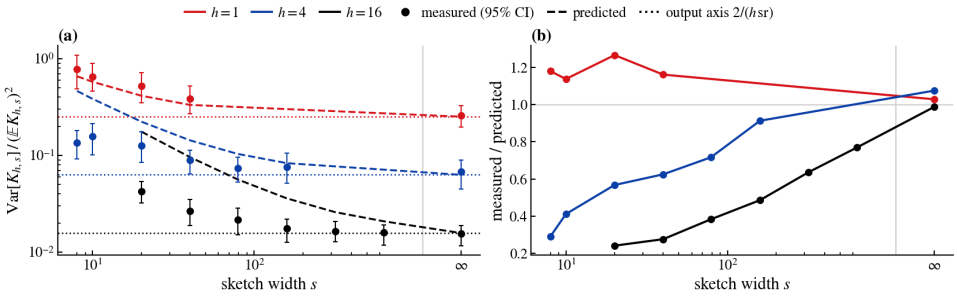

Figure 8: **Proposition 3.2 on a real eNTK block** ($r_{\mathrm{tr}}(\bar{M}) = 7.9$, $p = 1864$ unpadded parameters, $d = 8$, 96 draws per cell, 19 (h, s) cells). **(a)** Measured relative variance with 95% CI against a prediction that uses an h-independent JL surrogate for the parameter axis (dashed). Dotted lines mark the output term $2/(h\, r_{\mathrm{tr}})$. **(b)** Measured over predicted. The unsketched cells ($s = \infty$) give 1.03, 1.07 and 0.99. The sketched cells run from 1.27 at $h = 1$ to 0.24 at $h = 16$, because the surrogate lacks the $1/h$ term of Eq. 4.